METODOLOGÍA PARA LA CARACTERIZACIÓN DE IMÁGENES EN EL RECONOCIMIENTO DE LENGUAJE DE SEÑAS COLOMBIANO Y SU TRADUCCIÓN AL ESPAÑOL

Ing. Arley Bejarano Martínez

Proyecto de grado presentado como requisito parcial para aspirar al título de Magíster en Ingeniería Eléctrica

Director
Ing. Julian David Echeverry Correa. M.Sc. Ph.D.

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS ELÉCTRICA, ELECTRÓNICA, FÍSICA Y CIENCIAS DE LA COMPUTACIÓN
PEREIRA
2019

El presente trabajo de grado hace parte del proyecto “Metodología para el Reconocimiento y la Traducción de Señas Aisladas en la Lengua de Señas Colombiana Utilizando Técnicas de Visión por Computador” avalado por la Vicerrectoría de Investigaciones, Innovación y Extensión de la Universidad Tecnológica de Pereira con código 6-16-4. Presentado en la convocatoria interna de financiación de proyectos de investigación 2015 (programas académicos)

Dedicado a mi esposa Yuliana Sanchez Rendón y mis padres Juan Bejarano y Elsa Martínez quienes con su esfuerzo me ayudaron a culminar este proceso.

Agradecimientos. A mi esposa Yuliana Sánchez Rendón quien a través de sus consejos me ayudó a terminar con éxito esta etapa de mi vida. A mis padres Juan Bejarano y Elsa Martínez quienes me han ayudado con sus consejos, oraciones y ejemplo de vida a conseguir este logro. A mi director, el ingeniero Julian David Echeverry quien me ayudó en cada paso de este proyecto por sus consejos y recomendaciones. Y por último a mis amigos que compartieron sus conocimientos y me ayudaron a culminar este trabajo.

Índice

Índice de tablas

Índice de figuras

Lista de acrónimos

- BLEU: BiLingual Evaluation Understudy
- BMP: Bitmap - Mapa de bits.
- HMM: Hidden Markov Model - Modelos ocultos de Márkov
- IBM: International Business Machines
- INSOR: Instituto Nacional para Sordos
- LM: Language Model - Modelo de Lenguaje
- LSC: Lengua de Señas Colombiana
- MT: Machine Translation - Traducción Automática
- OMS: Organización Mundial de la Salud
- SDK: Software Development Kit - Conjunto de herramientas para el desarrollo de software
- SER: Sentence Error Rate - Tasa de Frases Erróneas
- SMT: Statistical Machine Translation - Traducción Automática Estadística
- TM: Translation Model - Modelo de Traducción
- TXT - Text
- UTP: Universidad Tecnológica de Pereira
- WER: Word Error Rate - Tasa de Palabras Erróneas

RESUMEN

El lenguaje de señas según la OMS es utilizado por 360 millones de personas en el mundo que presentan pérdida parcial o total de la audición. Por medio de este tipo de lenguaje logran expresar sus ideas y sentimientos. Este tipo de personas cuentan con un obstáculo a la hora de estudiar o realizar alguna labor ya que el mundo actual está diseñado para personas que se comunican por medio de lenguaje hablado. Debido a esto y a los avances tecnológicos, varios centros de investigación y universidades han desarrollado metodologías para traducir de lengua de señas a lenguajes hablados y viceversa, para esto se han utilizado sensores de electromiografía, masa inercial y cámaras. Con estos sensores se captura la información de movimiento y se extraen características mecánicas, como ángulos, velocidad, energía, aceleración, entre otros. Con la información obtenida por medio de estos sensores se utilizan técnicas de aprendizaje de máquina como redes neuronales artificiales, máquinas de vectores de soporte, árboles de decisiones, entre otros.

En este documento se muestra la implementación de una metodología para el reconocimiento de lengua de señas y su traducción al español. Para esto se creó una base de datos con seis personas, cuatro hombres y dos mujeres que presentan diferentes medidas antropométricas. Dentro de los actores que realizaron las señas, cinco son hablantes nativos y uno es intérprete. La información se capturó por medio de una cámara de profundidad y se obtuvo un total de 199 señas aisladas y 51 señas compuestas. Con la información obtenida se realizó la extracción de diferentes características, luego se agruparon y se validaron por medio del coeficiente silhouette. Los datos obtenidos en el agrupamiento son utilizados para entrenar diferentes modelos ocultos de Márkov. Para validar esta metodología se realizó una validación cruzada de seis partes, se repite el proceso seis veces cambiando la cantidad de observaciones de las HMM, se reportaron los porcentajes de acierto.

Una vez se identificaba la seña se procedió a implementar un sistema de traducción automático de glosas a español, para esto se cargo un sistema estadístico de traducción automático, el cual es entrenado con 1919 palabras y 517 frases. Estas frases se dividieron de forma aleatoria en 10 particiones y se realizó una validación cruzada obteniendo así diez modelos diferentes de traducción. Para cada uno de los modelos obtenidos se sintonizan los parámetros con el fin de mejorar la traducción. Luego se validó el sistema de traducción por medio de WER y BLEU.

ABSTRACT

Sign language according to the WHO is used by 360 million people in the world who have partial or total hearing loss. Through this type of language, they manage to express their ideas and feelings. This type of people has an obstacle when it comes to studying or doing some work since the current world is designed for people who communicate through spoken language. Due to this and the technological advances, several research centers and universities have developed methodologies to translate from sign language to spoken languages and vice versa, for this purpose electromyography sensors, inertial mass sensors and cameras have been used. With these sensors, the movement information is captured and the mechanical characteristics are extracted, such as angles, speed, energy, acceleration among others. With the information obtained through these sensors, machine learning techniques, such as artificial neural networks, support vector machines, decision trees, among others, are used. This document shows the implementation of a methodology for the recognition of sign language and its translation into Spanish. For this, a database with six people was created, four men and two women who have different anthropometric measurements. Among the actors who performed the signs, five are native speakers and one is an intérpreter. The information was captured by means of a depth camera and a total of 199 isolated signs and 51 composite signs were obtained. With the obtained information, the extraction of different characteristics is performed, then a grouping is carried out that is validated by means of the silhouette coefficient. The data obtained in the grouping are used to train different hidden Markov models. To validate this methodology, a cross-validation of six parts is performed, the process is repeated six times, changing the number of observations of the HMM, and a percentage of success is reported. Once the sign is identified, a system of automatic translation of glosses into Spanish is implemented, for which a statistical system of automatic translation is proposed, which is trained with 1919 words and 517 phrases. These sentences are split randomly into 10 partitions and a cross-validation is carried out, thus obtaining ten different translation models. For each of the models obtained, the parameters are tuned in order to improve the translation. In order to validate the translation system, the WER and BLEU were calculated.

1. PRELIMINARES

1.1. Introducción

Uno de los ejes fundamentales para cualquier sociedad es lograr conseguir la inclusión de todos los individuos como personas activas de la comunidad, esto es una labor compleja cuando se consideran a las personas que presentan discapacidades físicas o mentales. Dentro de las personas que presentan discapacidades se encuentran aquellas que presentan sordera parcial o total, las cuales corresponden a 360 millones en el mundo según la OMS. Este tipo de personas se comunica a través de la lengua de señas, para esto utilizan las manos y el cuerpo, y así expresan sus emociones y pensamientos con otros individuos. Este tipo de discapacidad en muchos casos genera un aislamiento de la persona, ya que comprender o interactuar con las personas que presentan sordera parcial o total se vuelve una tarea compleja, lo que genera una barrera a la hora de hacer tareas sencillas como estudiar, trabajar, entre otras labores cotidianas.

La lengua de señas es una lengua natural de expresión y configuración gesto-espacial y percepción visual. Cuenta con la característica de ser ágrafa, es decir, que no cuenta con escritura, lo que dificulta su interpretación. Para eliminar esta dificultad de interpretación se crearon las glosas que permiten la transcripción de las señas a un medio escrito representando así su concepto.

Dada la dificultad y la cantidad de personas que se comunican por medio de esta lengua se han desarrollado sistemas que permiten identificar la glosa que representa la seña, lo cual ha generado que se enfoquen en la extracción de características de señales capturadas por sensores de masa inercial, acelerómetros, o videos y su respectiva clasificación por medio de técnicas de aprendizaje de máquina. Por otro lado, se han desarrollado sistemas de traducción de glosas al español.

El presente trabajo de grado hace parte del proyecto "Metodología para el Reconocimiento y la Traducción de Señas Aisladas en la Lengua de Señas Colombiana Utilizando Técnicas de Visión por Computador" avalado por la Vicerrectoría de Investigaciones, Innovación y Extensión de la Universidad Tecnológica de Pereira. Este trabajo se enfoca en dos temáticas, la primera es el reconocimiento de las señas, donde se utilizó una cámara de profundidad para capturar la información, se realizó una extracción de características y su respectivo entrenamiento por medio de modelos ocultos de Márkov. La segunda es la implementación de un sistema estadístico de traducción automática de glosas a español; cada uno de los métodos realizados se valida por medio de validación cruzada y se reporta el porcentaje de acierto.

El fin último del proyecto de investigación es facilitar la comunicación entre personas sordas y personas oyentes con el fin de que las personas sordas encuentren distintas fuentes de información dentro de la Universidad Tecnológica de Pereira.

Dicho esto, esta memoria está organizada en siete capítulos repartidos de la siguiente manera: en los capítulos 1, 2 y 3 se presenta la introducción a este proyecto de grado, el planteamiento del problema y los objetivos que se cumplieron durante el desarrollo del trabajo. En el capítulo 4 se dan los fundamentos teóricos que respalda el desarrollo y la ejecución de este trabajo. En el capítulo 5 se explica detalladamente cada uno de los pasos realizados para el cumplimiento de los objetivos y los resultados obtenidos en cada uno de ellos. En los capítulos 6 y 7 a partir de los resultados obtenidos

se extraen las conclusiones y los posibles trabajos a futuro. Finalmente, las referencias donde puede consultarse toda la bibliografía utilizada.

1.2. Definición del problema

Una de las necesidades de los seres humanos como parte activa de una sociedad es comunicarse, ya que a través de la comunicación se transmite y se comparte información e ideas. Desde la creación de las primeras comunidades el ser humano ha intentado comunicarse, primero a través del lenguaje escrito utilizando jeroglíficos, símbolos o dibujos. Luego a través de sonidos, y posteriormente por medio del lenguaje hablado.

Con el pasar del tiempo se fueron creando modelos y herramientas que permitieron una mejor comunicación entre las personas que conformaban las comunidades. Así fue como el mundo vio nacer el alfabeto, el papel, los libros, los periódicos, el telégrafo, el teléfono y en la actualidad la radio, la televisión y el Internet [1].

A lo largo de la historia el hombre ha tenido la necesidad de comunicarse con otras comunidades con el fin de intercambiar ideas, experiencias, información, facilitar el comercio y solucionar problemas en común. A partir de esa necesidad se han creado diferentes formas de solucionar el problema de comunicación entre personas que manejan distintas lenguas maternas; la opción inicial consistía en que alguien aprendiera el lenguaje de ambas comunidades e hiciera la traducción entre los miembros de estas. Actualmente, esta opción no resuelve el problema de comunicación en un mundo globalizado que cuenta con más de siete mil idiomas diferentes [2].

En el siglo pasado, aprovechando el avance de la tecnología en el campo del procesamiento del lenguaje natural, y el auge del Internet, aparecieron los sistemas de traducción automática de texto, los cuales no solo facilitaron la tarea de traducción sino que permitieron automatizarla.

En los últimos años, y debido al aumento de dispositivos móviles se crearon los traductores simultáneos de voz [3, 4], los cuales se basan en el procesamiento del lenguaje hablado [5].

Más del 5 % de la población mundial sufre sordera parcial o total [6], gran parte de esta población se comunica utilizando la lengua de señas, este tipo de lenguaje es ágrafo lo cual dificulta que las personas que se comunican con la lengua de señas tengan problemas a la hora de comunicarse con las personas que utilizan lenguajes hablados. Esta dificultad genera que las personas que sufren de sordera no pueden integrarse en sus diferentes círculos sociales, al punto de sentirse aislados por su condición [7].

La lengua de señas tiene como elemento básico la seña *per se*, la cual se compone de una parte visual, una manual y una gestual; y cuenta también con reglas gramaticales y sintácticas al igual que los lenguajes hablados. Para interpretar este lenguaje se debe tener en cuenta la configuración, posición y orientación de cada una de las señas respecto al individuo y al espacio en que se ubica. Lo primero a tener en cuenta en la interpretación de este lenguaje es la ubicación y orientación de las manos y dedos, lo que se conoce como el estudio de la **matriz articulatoria**. Lo segundo son los rasgos no manuales, los cuales a su vez se dividen en dos tipos: el primero son las reglas morfológicas y corresponden al orden en que se hacen las señas; y el segundo son los movimientos y detenciones que conforman las señas. Al conjunto de estos dos tipos de rasgos se le da el nombre de **matriz segmental** [8].

Se puede apreciar que para determinar una seña es necesario conocer varios aspectos del movimiento que realiza la persona y la secuencia en que se hace la misma. Es por esto que se han diseñado sistemas de captura de movimiento basados en acelerómetros, sensores de masa inercial y sensores de electromiografía que han permitido capturar y traducir de la lengua de señas a otros lenguajes.

Dentro de este tipo de sistemas se encuentra un desarrollo hecho en el Instituto Tecnológico de Rochester, esta investigación tiene como sistema de adquisición la banda Myo. Se realizó la captura de 26 letras del alfabeto inglés y se utilizan descriptores estadísticos, como media, desviación, máximos y mínimos a las señales entregadas por los sensores de electromiografía. Estos descriptores son utilizados para entrenar una máquina de soporte vectorial obteniendo una eficiencia del 80 % en la clasificación [9].

Otro de los desarrollos que utilizan sensores de masa inercial y sensores de electromiografía se desarrolló en el Instituto Tecnológico de Karlsruhe. En este trabajo se realizó la captura de 12 movimientos, se calcula la energía de la señal y la desviación estándar con el fin de extraer las características e hicieron el entrenamiento de una HMM obteniendo un 97.8 % de eficiencia en la clasificación [10].

Aunque estos sistemas presentan resultados favorables, tienen el problema de contar con sistemas de adquisición invasivos, lo que dificulta que se generen movimientos naturales, además, que si se desea escalar el sistema es necesario crear una base de datos que contenga cada una de las señas lo cual generaría un problema, debido a que se contaría con una gran cantidad de información.

Para solucionar el problema de analizar y clasificar grandes flujos de datos, se planteó la metodología utilizada para traducción simultánea de voz. Esta metodología realiza un reconocimiento de fonemas, en vez de reconocer frases completas, ya que computacionalmente es una estrategia eficiente, debido a que existen muchos menos fonemas que frases. Para la identificación de los fonemas se realiza la transformada Cepstrum y se calculan los coeficientes Mel, con esta información se utilizan modelos de clasificación como KNN o K-means, los cuales se encargan del reconocimiento de fonema, con estos fonemas se realiza el reconocimiento de palabras y frases utilizando alineamiento temporal o modelos ocultos de Márkov pre-entrenados [11].

Adicional a la reducción de datos se han implementado sistemas de captura no invasivos por medio de cámaras de profundidad. Dentro de este tipo de investigaciones la escuela de ingeniería de Amrita en la India [12] desarrolló un sistema para el reconocimiento de señas del lenguaje hindú. En este caso se realizó el estudio a señas aisladas sin movimiento, se utilizó OpenCV y se extrajeron características de distancia entre los puntos de interés obteniendo una eficiencia promedio del 40 %. Otro desarrollo se presenta en la Universidad Tecnológica de Rzeszow donde se realizó un sistema de traducción automática de lengua de señas polaco al idioma polaco, para este estudio se realizaron grabaciones de 30 señas con 10 intérpretes, con esta información se implementó un algoritmo de clasificación por medio de vecinos más cercanos y alineación temporal en tiempo, obteniendo un 98.33 % en la clasificación [13]. En el Instituto de Tecnología de Jaypee en la India, se realizó una investigación en la cual se realiza una preclasificación de las diferentes señas en pequeños movimientos utilizando K-means; para esta labor se utilizan vectores normalizados de los puntos articulados del cuerpo. Los centroides entregados por K-means se vuelven la entrada del HMM, con esto se realizó una primera fase con el fin de tener una aproximación de la palabra que se realizó, luego se calcula la distancia entre la palma de la mano y los dedos con el fin de obtener mayor información y así realizar una segunda clasificación para determinar la seña realizada. El porcentaje de acierto de la traducción fue de 90.6 % [14].

En Colombia se han planteado varios métodos para darle solución a esta problemática; en la Universidad de San Gil crearon un sistema para la traducción de lengua de señas a texto, la base de datos construida consta de grabaciones de dos personas que realizaron cada una un total de nueve señas. Este prototipo utilizó la biblioteca de OpenCV y PCL (Point Cloud Library), se aplica la técnica de análisis de componentes principales (PCA) y para la clasificación se utilizan redes neuronales artificiales obteniendo un porcentaje de acierto de 94.2 % [15]. En la Universidad de los Andes se realizó una investigación en la cual se utilizaron seis personas no hablantes, se capturaron 17 señas estáticas y se utilizó una máquina de soporte vectorial para identificarlas; 4 señas aisladas con movimiento las cuales se identifican por medio de HMM. Para la identificación de las 6 palabras se realizó el mismo proceso, los descriptores utilizados para extraer las características fueron los descriptores invariantes de Fourier, delta de movimiento y cosenos directos, obteniendo un acierto de 94.7 % en letras estáticas, 93.8 % en letras con movimiento y 99.1 % en palabras [16].

El problema es que según las investigaciones encontradas, cuando se realizó la consulta del estado del arte, demuestran que se ha trabajado en mayor medida con bases de datos pequeñas de movimientos específicos o de algunas frases, y que no permiten extrapolar la idea de estos estudios ya que no se evidencia el estudio como un lenguaje sino que se enfocan en los métodos de clasificación y de extracción de características para las manos o rasgos particulares. Por lo tanto, no se han definido las formas mínimas o fonemas articulados que conforman las señas y así realizar de manera más sistémica el estudio de este lenguaje, como se hace con los lenguajes hablados, tampoco se cuenta con una base de datos pública de lengua de señas colombiana que sea amplia y que contenga tanto señas aisladas como compuestas [9, 10, 12, 14, 15, 13, 16].

Para intentar solucionar el problema ya mencionado se creó una base de datos que contiene 199 señas aisladas y 51 señas compuestas de la lengua de señas colombiana que realizaron cinco personas no oyentes y un intérprete de lengua de señas sobre un dominio específico como es el centro de información de un centro educativo. La información se capturó por medio de una cámara de profundidad con el fin obtener la información del movimiento que realiza la persona. Con esta información se aplicaron diferentes algoritmos del estado del arte y propios con el fin de realizar la extracción de características mecánicas y se realizó una primera clasificación no supervisada por medio de K-means con el fin de determinar los fonemas articulados, para validar el agrupamiento se utiliza el coeficiente de silhouette; una vez definido los fonemas articulados se realizó el entrenamiento de diferentes modelos oculto de Márkov, todo esto con el fin de determinar la seña aislada o compuesta que fue realizada, la información obtenida es pasada un modelo de traducción estadístico para luego ser evaluado por medio de BLEU y WER.

Este proyecto surge con la motivación de avanzar hacia la caracterización de la lengua de señas colombiana utilizando técnicas de visión por computador y planteando una metodología para la identificación de las glosas, con el fin de analizar la lengua de señas de una manera sistémica y no solo desde la parte de la caracterización, es por esto que este documento presenta dos grandes temáticas. La primera es la caracterización de las señas y el reconocimiento de la misma utilizando técnicas de aprendizaje de máquina, la segunda parte se centra en realizar un sistema de traducción automático estadístico, que permita pasar de glosas a español. Con este proyecto se podría ayudar a la comunidad sorda, facilitando la forma de interactuar ante personas oyentes, también se podría crear una mayor

participación de este tipo de personas en las instituciones de educación superior mejorando la calidad de vida de estas personas.

1.3. Objetivos

1.3.1. Objetivo general

Desarrollar y evaluar una metodología para la caracterización de imágenes en el reconocimiento de lenguaje de señas colombiano y su traducción al español.

1.3.2. Objetivos específicos

- Crear una base de datos anotada compuesta por imágenes y videos de intérpretes de lenguaje de señas colombiano en un dominio de aplicación específico. Esta base de datos tendrá realizaciones propias de un sistema de información en una institución de educación superior.
- Determinar un espacio vectorial de representación para las características que se obtienen a partir de los puntos articulados entregados por un sensor de profundidad.
- fórmular una metodología para la creación de un diccionario de «fonemas» articulados utilizando técnicas de agrupamiento supervisado.
- Determinar una metodología para el reconocimiento automático de la lengua de señas utilizando técnicas de aprendizaje de máquina que hagan uso del diccionario de fonemas articulados creado para este trabajo.
- Validar estadísticamente el rendimiento de las metodologías planteadas mediante métricas empleadas en el campo del procesamiento del lenguaje natural tales como BLEU, WER.

2. ESTADO DEL ARTE

En la sección 2.1 de este capítulo se mostrarán las bases de datos disponibles que contienen diferentes lenguas de señas, los avances utilizados hasta el momento a nivel mundial para la identificación de señas aparecen en la sección 2.2 y los sistemas de traducción automáticos en la sección 2.3, allí se da un contexto sobre los elementos y técnicas utilizadas para realizar la labor de traducir señas a algún lenguaje hablado o escrito.

2.1. Bases de datos

Uno de los puntos claves de este trabajo de grado es la creación de la base de datos, esto se debe a que hasta lo consultado a la fecha no existen base de datos públicas de la lengua de señas colombiana.

Las bases de datos que se encuentran, se pueden clasificar en dos, las primeras que tienen como fin enseñar la lengua de señas de diferentes países presentando señas en videos e imágenes con el fin de generar una sociedad incluyente, algunas de ellas son:

- La ASL-LEX la cual cuenta con alrededor de 1000 señas disponibles en video, presenta señas aisladas como señas compuestas [17].
- La ASLRRP consta de 9800 realizaciones las cuales fueron realizadas por seis sordos nativos y presentan señas de la lengua de señas estadounidense [18].
- La BSL SignBank cuenta con 50000 realizaciones de 2500 señas de la lengua de señas británico [19].
- ASL SignBank, esta fue desarrollada por la universidad de Yale y cuenta con 664000 realizaciones de personas sordas entre niños y mayores [20].
- Asian SignBank creada por el centro de estudio de lingüística y sordos de Asía, presenta imágenes y videos de las diferentes lenguas de signo del continente asiático [21] .
- Auslan SignBank adquirida con la universidad de Cambridge, presenta 6500 realizaciones de la lengua de señas australiana [22].

Las segundas bases de datos que se encuentran, son las que se utilizan en investigaciones de traducción y compresión de las diferentes lenguas de señas, dentro de estas se encuentran:

- La DGS Kinect 40, la cual contiene 40 signos del lenguaje de señas alemán y es realizada por 15 personas, cada seña tiene 3000 repeticiones, se almacenan videos y múltiples ángulos de las personas que realizan las señas [23].
- RWTH-PHOENIX-Weather. Presenta 1200 señas realizadas por 9 personas, cada seña cuenta con 45760 repeticiones entre señas aisladas y compuestas del lenguaje de señas alemán [24].
- La base de datos SIGNUM presenta señas del lenguaje alemán, contiene 450 señas realizadas por 25 personas para un total de 33210 muestras que contienen señas aisladas y compuestas [25].

- GSL 20 es una base de datos que presenta signos del lenguaje de señas Griego, presentando 20 señas diferentes que realizaron 6 personas, generando 840 datos almacenados en videos que contienen señas asiladas [26].
- La base de datos PSL ToF 84 cuenta con 1680 muestras de señas aisladas polacas, almacenadas en videos capturados con una cámara de profundidad [27].
- PSL Kinect 30 es una base de datos de lenguaje de señas polaco que contiene 30 señas realizadas por una persona, generando un total de 300 muestras de señas aisladas, la captura se realiza por medio del kinect y se almacenaron los videos de profundidad [28].
- La base de datos Boston ASL LVD cuenta con 3300 señas realizadas por 6 personas se almacenaron 9800 videos y ángulos de signos aislados del lenguaje de señas estadounidense [29].
- MSR Gesture 3D es una base de datos que contiene signos aislados del lenguaje de señas estadounidense que se representan en 12 clases realizadas por 10 personas, generando 336 videos capturados con una cámara de profundidad [29].
- Una de las base de datos más completa fue capturada en China por el grupo VIPL [30] para el centro de investigación y desarrollo de la empresa Microsoft. Esta base de datos fue creada por 8 personas y presenta tanto señas aisladas o compuestas, la organización de esta base de datos es la siguiente:
  - DEVISIGN-G Presenta 36 clases de 8 personas, cada una de 432 muestras con videos que contienen palabras.
  - DEVISIGN-D Presenta 500 clases de 8 personas, cada una de 6000 muestras con videos que contienen palabras.
  - DEVISIGN-L Presenta 2000 clases de 8 personas, cada una de 24000 muestras con videos que contienen palabras.
- IIITA –ROBITA es una base de datos que contiene videos de 23 diferentes signos aislados de la lengua de señas hindú [31].
- LSA64 contiene 3200 videos de signos aislados de la lengua de señas Argentina, para realizar esta base de datos se utilizaron 10 personas y se definieron 64 clases [32].
- LIBRAS, se basa en lenguaje de señas brasilero consta de 610 realizaciones que contienen imágenes y trayectorias capturadas por medio del kinect [33].

2.2. Reconocimiento de señas

Dado que para implementar un sistema automático de traducción es necesario reconocer una seña se hace indispensable conocer los movimientos que realiza la persona como se explica en la sección 3.1, se han planteado y estudiado principalmente el uso de dos tecnologías para el reconocimiento de las mismas, la primera utiliza sensores invasivos (IMU, acelerómetros, EMG) y otro enfoque que captura la información por medio de sensores no invasivos (cámaras a color, cámaras de profundidad, cámaras de tiempo de vuelo, entre otras).

2.2.1. Sensores invasivos

Los primeros sistemas de traducción de lenguaje de señas a lenguaje hablado tuvieron como base los sensores de electromiografía y sensores de masa inercial [10, 9, 34, 35].

En la escuela de computación de la Universidad de Colombo realizaron una investigación con el sensor Myo que contiene acelerómetros, giroscopios y sensores de electromiografía, para la creación de la base de datos se realizó la captura de 5 señas con 6 personas la mitad mujeres y la otra mitad hombres. Con la información obtenida se calcularon la media y la desviación estándar de la energía y con estas características se realizó el entrenamiento de una red neuronal artificial arrojando resultados de acierto entre 77.5 % y 97.6 % [34]. Otro desarrollo, se realizó en la Universidad A&M de Texas, en este caso utilizaron fusión de datos con IMU y sensores de electromiografía, se creó una base de datos que consta de 40 señas aisladas del lenguaje de señas estadounidense, en este estudio se plantean diferentes descriptores y clasificadores con el fin de demostrar el mejor desempeño. Las características que fueron extraídas fueron medias de los sensores, calculó de valores RMS del movimiento, varianza de las medidas, entre otros. Por otro lado para el proceso de clasificación utilizaron redes Bayesianas, alineamiento temporal, redes neuronales artificiales y máquinas de vectores de soporte, el mejor desempeño fue con el método de SVM y los datos obtenidos con los sensores de electromiografía [35]. En el Instituto de Ciencia y Tecnológica de Deagu en Corea, allí realizaron la captura de 30 gestos de la lengua de señas coreana por medio de la MYO y se realizó un entrenamiento de una red neuronal convolucionada obteniendo resultados de acierto de 97.12 % [36]. En todos estos desarrollos la captura de los gestos se realizó de manera aislada, es decir, que no se realizaron captura a nivel de frases completas sino a nivel de glosas aisladas, por lo tanto, no se consideró un diálogo fluido. Para este trabajo el diálogo fluido se definió como seña completa.

2.2.2. Sensores no invasivos

En el centro para la Visión, el Habla y el Procesamiento de Señales de la Universidad de Surrey ubicada en el Reino Unido se desarrolló uno de los primeros estudios para la traducción de lenguaje de señas utilizando cámaras de profundidad. En este se capturaron las señas del alfabeto americano de lenguaje de señas y se realizó una base de datos con cuatro personas. La principal característica de este desarrolló fue identificar los dedos, por lo tanto plantean el algoritmo de HOG (Histograma de gradiente orientado) el cual permite la identificación de objetos y formas dentro de una imagen y se realiza un entrenamiento utilizando HMM, obteniendo una tasa promedio acierto de 50.84 % [37]. Cabe resaltar que en esta investigación las señas eran estáticas y el enfoque principal de este trabajo se basó en la extracción de características de las manos.

En la Universidad de Sharjah ubicada en Arabia Saudita se planteó una metodología para el reconocimiento de señas aisladas del lenguaje de señas arábigo utilizando el discriminante lineal de Fisher. En este estudio se realizaron videos a 3 personas diferentes realizando 23 señas, luego se extrajeron características de movimiento del video y se aplica la transformada discreta del coseno. A estas características se les aplicó el discriminante lineal de Fisher para clasificar las señas obteniendo un 61.6 % de acierto [38].

En Arabia Saudita en la Universidad de Minerales y Petróleo del Rey Fahd se realizó un estudio para el reconocimiento de lenguaje de señas arábigo. En este se realizaron la captura de 300 videos,

a cada uno de ellos se les extrajo la posición de las manos utilizando técnicas de segmentación y seguimiento en el espacio de color HSI y se realizó la clasificación por medio de una HMM obteniendo un rendimiento del 93 % [39].

Dentro de estos desarrollos se encuentra en la Escuela de Ingeniería de Amrita en la India un sistema para el reconocimiento de señas del lenguaje hindú. En este caso se realizó el estudio de señas aisladas sin movimiento. Se utilizó OpenCV y se extraen características de distancia entre los puntos de interés obteniendo un acierto promedio del 40 % [12].

Otro desarrollo se presenta en la Universidad Tecnológica de Rzeszow ubicada en Polonia, donde se realizó un sistema de traducción automática del lenguaje de señas polaco a idioma polaco. Se grabaron 30 palabras con 10 intérpretes. Se realizó una clasificación por medio de vecinos más cercanos y alineación temporal en tiempo, obteniendo un 98.33 % de acierto [13]. Este desarrollo presenta una base de datos con pocas palabras.

En el Instituto de Tecnología de Jaypee en la India, se realizó una investigación en la cual se realizó una preclasificación de la seña en pequeños movimientos utilizando K-means. Para esta labor se realizaron dos etapas. Para la primera utilizaron vectores normalizados de los puntos articulados del cuerpo. Con estos vectores se realizó un entrenamiento no supervisado utilizando K-means, el conjunto de datos entregados por el entrenamiento fueron utilizados como observaciones del modelo oculto de Márkov, con esto se redujeron la cantidad de clases a detectar. Para la segunda etapa se realizó la extracción de la información de la palma de la mano y de los dedos calculando la distancia entre ellos y así con esta información se realizó una segunda clasificación con el fin de aumentar el porcentaje de acierto de la traducción obteniendo 90.6 % [14].

Las anteriores investigaciones se enfocan en la extracción de características de las manos o señas, sin embargo no realizan el análisis del lenguaje de señas desde el punto de vista de las partes que conforman una seña. Una de las investigaciones que pretende atacar este problema se plantea en el Departamento de Ingeniería Electrónica y Ciencias de la Información de la Universidad de Ciencia y Tecnología de China quienes plantearon una metodología para la caracterización de señas en fotogramas claves. Para detectar estos fotogramas claves utilizaron una fórmula que determina la cantidad de movimiento realizado entre las diferentes partes de la seña. Para comprobar la metodología propuesta crearon una base de datos que contiene 310 video de 100 señas realizadas por 10 personas las cuales fueron capturadas por medio del sensor Kinect 2 y consta de señas del lenguaje de señas chino. En cada uno de estos puntos claves se extrajeron características de la cámara RGB, la cámara de profundidad, de los puntos articulados, y se aplica HOG. Con estas características se realiza la comparación de diferentes métodos de clasificación como convolución de redes neuronales artificiales, alineamiento temporal y HMM obteniendo aciertos entre 86.12 % y 91.18 % [40].

En Colombia, en la Universidad de los Andes se realizó una investigación en la cual se utilizaron seis personas sordas que realizaron videos de 17 señas estáticas aisladas, se realizó la clasificación utilizando una máquina de soporte vectorial, también se capturaron 4 señas aisladas con movimiento las cuales se identifican por medio de modelos ocultos de Márkov y 6 palabras utilizando nuevamente HMM, el sistema de adquisición utilizado fue el sensor Kinect y para extraer la información relevante se usaron los descriptores invariantes de Fourier, delta de movimiento y cosenos directos, obteniendo un acierto de 94.7 % en letras estáticas, 93.8 % en letras con movimiento y 99.1 % en palabras [16].

Otro desarrollo en Colombia se presenta en la Universidad de San Gil, donde crearon un sistema para la traducción de lenguaje de señas a texto, la base de datos construida consta de dos personas que realizaron cada una un total de nueve señas, este prototipo utilizó OpenCV y PCL (Point Cloud Library), se utilizó análisis de componentes principales y se utilizaron redes neuronales artificiales para la clasificación obteniendo un porcentaje de acierto de 94.2 % [15].

2.3. Sistemas de traducción automática

Los sistemas de traducción automática son sistemas que se fundamentan en técnicas de aprendizaje de máquina y la lingüística, esta disciplina se encarga de comprender la forma en que se comunican las personas y así poder crear sistemas de traducción simultánea, sistemas de diálogo interactivos, análisis de opiniones, entre otros. Los primeros desarrollos de traducción automática aparecen en el año 1933, cuando el francés George Artsrouni y el ruso Petr Trojanski patentaron sus trabajos. El primero fue acerca del diseño de un dispositivo de traducción y el segundo, fue una propuesta de un método para un diccionario bilingüe automático. Con la creación de las computadoras usadas en Gran Bretaña para romper el código Enigma alemán en la Segunda Guerra Mundial y en la década de los 60 se conoció el primer traductor automático que traducía de ruso a inglés [41].

Después de estas propuestas de traductores automáticos, se empezaron a estudiar diferentes enfoques para conseguir sistemas de traducción automática, desde métodos de traducción directa, métodos de transferencia, hasta métodos de interlengua. En 1956, la Universidad de Georgetown y la compañía IBM realizaron el experimento Georgetown-IBM, que consistió en la traducción de ruso a inglés con un léxico que comprendía alrededor de sesenta frases. El experimento fue considerado un éxito, abriendo un camino de investigación hacia el área de la lingüística computacional. En Estados Unidos la mayoría de las investigaciones se centraban en la traducción del ruso al inglés por motivos políticos, mientras que, en Europa y Canadá, los estudios se enfocaban en la traducción inglés a francés. En 1966, el Comité Asesor para el Procesamiento Automático del Lenguaje (Automatic Language Processing Advisory Committee-ALPAC), publicó un informe donde afirmó que la traducción automática era costosa y que los intérpretes humanos abundaban, reduciendo así el interés en los desarrollos orientados a la traducción automática. Cuatro años después se crea el sistema Systran con el fin de realizar traducciones de ruso a inglés. En 1982, surgió el sistema Météo el cual traducía de forma automática pronósticos del tiempo. En la década de los ochenta, se comercializaron los sistemas Logos (traductor de alemán a inglés y de inglés a francés), Metal (traductor de alemán a inglés) y los sistemas desarrollados en la Organización Panamericana de la Salud (traductor de español a inglés e inglés a español). En la mayoría de los casos, los sistemas de traducción de la época, eran bajo los modelos clásicos de traducción (traducción directa, traducción por transferencia y traducción interlengua). En 1991, IBM publicó los resultados de los experimentos de un sistema denominado Candide. Este sistema se basaba en métodos exclusivamente estadísticos. El sistema Candide lo entrenaron con el corpus Hansard de las Actas del Parlamento Canadiense compuesta por aproximadamente tres millones de oraciones en inglés y francés. Actualmente, los sistemas de traducción automáticos han generado facilidades para la relación con otras culturas y la difusión del conocimiento. Muchos ejemplos claros son la traducción de páginas web en Internet, los traductores de un idioma a otro y las aplicaciones móviles. Otro ejemplo de las aplicaciones de la traducción automática es la traducción de las sesiones del Parlamento de

la Unión Europea. En estas sesiones se deben redactar actas en las 23 lenguas oficiales de la Unión Europea que dan pie a más de 506 combinaciones lingüísticas, ya que cada lengua puede traducirse a las otras 22 [41].

En la Universidad Tecnológica de Pereira se desarrolló un sistema de traducción automática basado en modelos estadísticos para la traducción de la lengua de señas colombiana al español, en este se plantea un método para la traducción de glosas al español, sin embargo no se tiene en cuenta la caracterización de la seña sino que se enfoca en la traducción de la misma, para evaluar el sistema se planteó BLEU y WER obteniendo valores de 28.53 % y 9.09 % respectivamente [41].

3. MARCO TEÓRICO

En este capítulo se encuentran los conceptos teóricos utilizados para desarrollar el trabajo propuesto.

En la sección 3.1 se dan a conocer los modelos propuestos para la interpretación del lenguaje de señas y la historia de cómo llega este tipo de lengua a Colombia.

Luego en la sección 3.2 se escribe sobre técnicas de agrupamiento o de clasificación no supervisada y se muestran las diferentes métricas para evaluar los agrupamientos. En la sección 3.3 se mencionan los modelos ocultos de Márkov los cuales se utilizarán más adelante para realizar la clasificación de las señas. Por último en la sección 3.4 se mencionan los sistemas de traducción automática y las métricas utilizadas para validar estos modelos

3.1. Lenguaje de señas

Debido a que en este trabajo se busca plantear una metodología que permita reconocer el lenguaje de señas colombiano, es necesario conocer las investigaciones que se han realizado con este tipo de lenguaje. El primero en desarrollar una investigación de este tipo fue el lingüista y profesor estadounidense William Stokoe demostrando que el lenguaje de señas americano es una lengua ya que posee gramática y una semántica propia. Con esta base se partió el estudio de este lenguaje como un lenguaje natural ya que se puede estudiar a nivel lingüístico (fonológico, morfológico, semántico y pragmático). Este tipo de lengua no es universal, ya que se presentan diferencias idiolectales y diatópicas entre las regiones o los países. Además, no dependen de otros sistemas de comunicación ni son iguales a los códigos gestuales usados por las personas de una cultura dada.

El modelo de lenguaje de señas propuesto hasta el momento que mejor lo describe es el propuesto por Liddell y R.E Johnson en el año 1989 [42]. Este modelo se basa en los siguientes aspectos:

- Indicaciones de los procesos de cambio que sufrían las formas segmentales subyacentes de las señas al manifestarse en superficie.
- Características segmentales de la morfología de las señas (los tipos de variación sufridos por varias señas debido a procesos morfológicos, como marcas de aspectos, cuantificadores, entre otros).

Por medio de este esquema, la estructura de cada seña se dispone a partir de un análisis de segmentos sucesivos en el tiempo, cada uno de los cuales recibe una representación individual que discrimina tres componentes principales:

- Matriz articulatoria: Está relaciona la posición de las partes móviles de las manos, su ubicación y su orientación. Estos rasgos se denominan rasgos articulatorios, dentro de esta se encuentra a su vez tres componentes, cada uno de los cuales comprende grupos de rasgos:
    - Configuración manual (CM): como se colocan los dedos y el pulgar.
    - Ubicación (UB): donde se ubica el articulador manual.
    - Orientación (OR): como está orientado el articulador.

- Matriz segmental: Define la actividad que realiza la mano y establece la existencia de tres tipos de segmentos, definidos por la longitud de períodos en los cuales cambian o no rasgos de la matriz articulatoria. Comprende el conjunto de rasgos que informan acerca de los tipos de segmento en los cuales se puede analizar la seña y de las características de la acción desarrollada en cada caso. Estos tipos de segmentos son:
  - Movimiento (M): período en el cual algún aspecto de la articulación está en cambio.
  - Detención (D): período en el cual ningún aspecto de la articulación cambia.
  - Transición (T): segmento de menor duración que una detención, pero con rasgos equivalentes a una de ellas.
- Actividad no manual: encierra los datos sobre la actividad significativa de los articuladores de la cara (boca, cejas, ojos), los movimientos de la cabeza y el cuerpo.

En este trabajo se busca modelar la matriz articulatoria y segmental del lenguaje de señas colombiano para así poder imitar el comportamiento de las personas que se comunican utilizando este tipo de lenguaje.

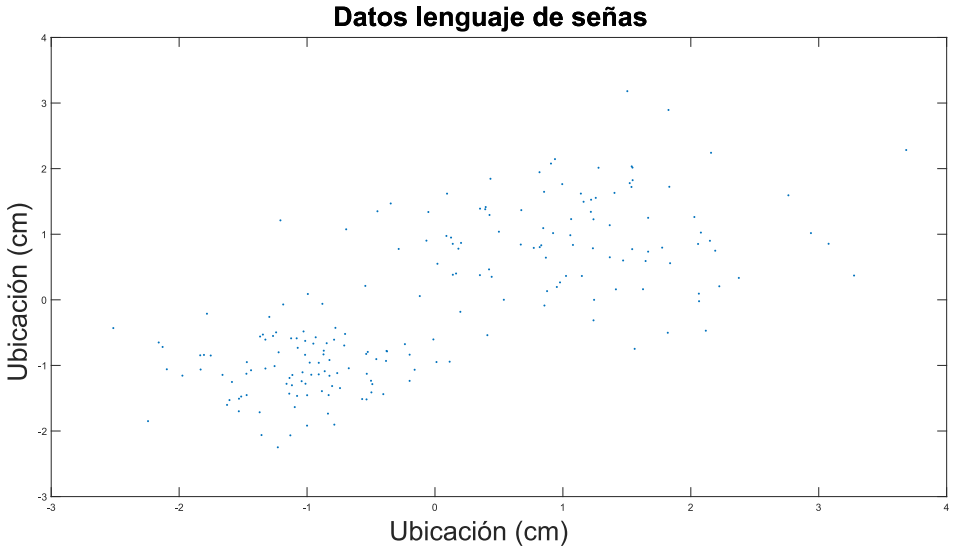
3.2. Técnicas de agrupamiento

Las técnicas de agrupamiento o *clustering* son algoritmos de reconocimiento de patrones no supervisados, este tipos de técnicas son altamente sensibles a los datos de entrada, ya que si se cuentan con una gran cantidad de datos atípicos, se presentara un sesgo en la clasificación y no podrá determinar de manera correcta la correspondencia de los datos. Dentro de estas técnicas se encuentra K-means, *spectral clustering*, entre otras. Estos métodos buscan características dentro de un conjunto de datos con el fin de crear subconjuntos que presentan información similar y así poder generar una clasificación.

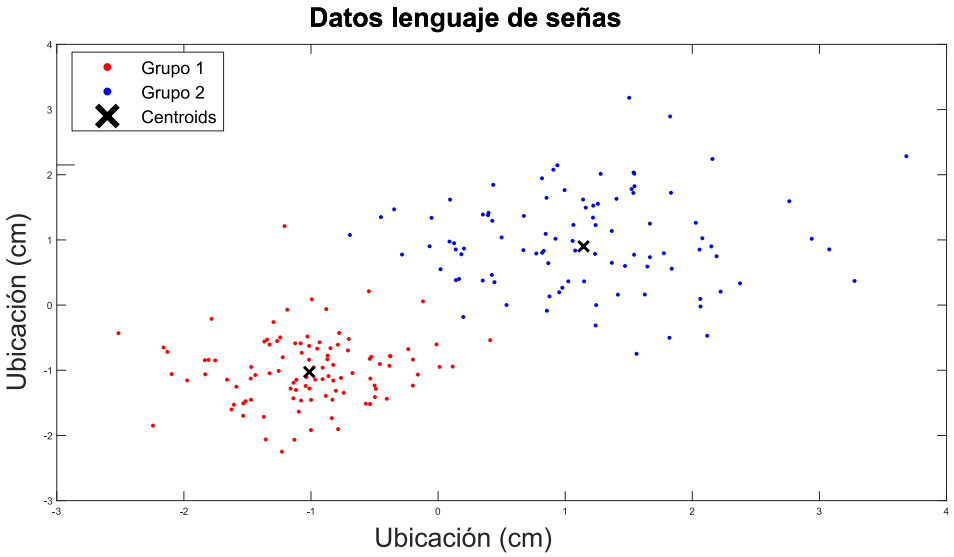
3.2.1. K-means

Creado por MacQueen en 1967 [43], es uno de los métodos de entrenamiento no supervisado más usados para la clasificación de un conjunto de objetos en un número K de grupos. Cada grupo está definido por centroides separados por una distancia euclidiana, como se observa en la figura 1[1]. En esta figura, en la parte 1a se muestra un ejemplo de los datos originales en un espacio bidimensional de dos características. Cada punto en la figura representa un frame de una de las señas que conforman la base de datos. En la parte 1b se muestra el resultado del agrupamiento de las señas en un número K de clusters, donde K=2 en este caso.

[1] Las figuras de este documento son propias a menos que se mencione lo contrario



(a) datos originales en un espacio bidimensional de dos características



(b) Resultado del agrupamiento de las señas en un número K de *clusters*

Figura 1: Separación de datos por medio de K-means

Dado un conjunto de observaciones $y = y_1, y_2, ...y_i$, el método de K-means tiene como objetivo agrupar j observaciones en $k \leq j$ grupos en los que cada observación pertenece al centroide del grupos más cercano, para la función objetivo P:

$$P = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - C_j||^2$$

Donde :

- k es el número de grupos

- n es el número de clases
- x_i es el dato a analizar
- Cj es el centroide del grupo j
- P es la función objetivo
- $||x_i^{(j)} - C_j||^2$ es la distancia euclidiana entre un punto y un centroide.

El algoritmo para determinar K-means es el siguiente

Algoritmo 1 Algoritmo de K-means

1. *a*) Elegir k puntos utilizando una distribución normalizada a partir de los datos con el fin de definir los centroides.
   b) Usando la distancia euclidiana calcular y almacenar las distancias entre cada k-centroides.
   c) Basado en la distancia calculada, a cada punto se le asigna el grupos más cercano.
   d) Calcular el promedio de las observaciones en cada grupo para obtener k nuevas ubicaciones de los centroides.
   e) Repetir los pasos de 2 a 4 hasta que las asignaciones de los gruposs no cambien o se alcance un número máximo de iteraciones.

3.2.2. Métricas para evaluación de grupos

Para determinar qué también agrupados se encuentran los grupos obtenidos, se han creado coeficientes o métricas que permiten medir la cohesión entre los datos y la separación entre los diferentes grupos, dentro de estos coeficientes se encuentra silhouette, Calinski y Harabasz, Hartigan, entre otros.

Coeficiente de silhouette

El coeficiente de silhouette es una métrica utilizada para medir la calidad de los resultados en algoritmos de agrupamiento. Por medio de este coeficiente se puede determinar la cohesión (que tan cercanos están los datos de un mismo grupo) y la separación (distancia entre los agrupamientos) [44].

El coeficiente de silhouette está definido para un punto x como:

$$S(x) = \frac{b(x) - a(x)}{max(a(x), b(x))}$$

Donde:

- $a(x)$ la distancia media del dato x que pertenece al cluster a.
- $b(x)$ la distancia media del dato x con los grupos b.

El valor de $S(x)$ varía entre -1 y 1. Valores cercanos a 1 demuestran una buena cohesión entre el dato analizado y su respectivo grupo y una buena separación entre el dato analizado y los demás grupos. Por el contrario un valor cercano a -1 quiere decir que hay una baja cohesión y una mala separación de

ese dato respecto al grupo que pertenece y los demás grupos. Los valores cercanos a 0 indican traslape entre los grupos.

Para conocer el coeficiente de todos el agrupamiento se calcula la media de cada uno de los coeficientes de silhouette obtenidos para todos los datos:

$$SC = \frac{1}{N} \sum_{i=1}^{N} S(x)$$

Donde N es la cantidad total de datos analizados.

3.3. Modelos ocultos de Márkov

Desarrollado por Baum a finales de los años sesenta, un proceso de Márkov es un proceso estocástico que sirve para representar secuencias de variables aleatorias independientes entre sí. Los Modelos Ocultos de Márkov son una herramienta estadística empleada en el modelamiento de series de tiempo, por lo que es un modelo doblemente estocástico el cual tiene un proceso oculto (secuencia de estados finitos) que sólo puede ser asociado probabilísticamente con otro proceso estocástico observable, produciendo la secuencia de características que se pueden observar. El principal objetivo de los HMM es encontrar el modelo que mejor explica una secuencia de observaciones dadas, dentro de un conjunto de modelos.

Una HMM está definido por:

- Los estados ocultos $h_i \ \varepsilon \ H$ con $i = 1, 2, ..., m$ donde m define el tamaño del vector de estados ocultos Ht.
- El vector de observaciones $X = x1, x2, ..., xn$,donde n define la longitud del vector de observaciones teniendo en cuenta que $m \leq n$.
- La distribución de los estados iniciales $\pi = (\pi_1, \pi_2, ..., \pi_n)$, donde el primer estado inicial h_0 es la probabilidad marginal (variable independiente e idénticamente distribuida). Por lo tanto:

$$[\pi] = p(h_i)$$

Para realizar la transición entre los estados ocultos h_i de una HMM, se necesita una matriz de t probabilidad de transición de estado $T \ \delta \ R^{m\,x\,m}$la cual viene dada por:

$$[T]_{j,i} = p(h_j|h_i)$$

Además se tiene la matriz de probabilidad de salida O que representa la distribución de probabilidad para muestrear un símbolo observable desde un estado oculto, la cual está definida como:

$$[O]_{k,j} = p(x_k|h_j)$$

Teniendo en cuenta las anteriores ecuaciones se construye λ, que es un arreglo donde el estado oculto del instante t depende únicamente del último estado oculto, por lo que una HMM se define como un 3-tupla $\lambda = (T, O, \pi)$ [45]. Por lo tanto:

$$\forall t : p(h_t|h_{1:t-1} = p(h_t|h_{t-1})$$

$$\forall t : p(x_t|h_t)$$

Donde $o_j(x_k)$ es la probabilidad de emitir un símbolo de observación en el estado h_j

3.4. Sistemas de traducción automática

Los sistemas de traducción automática conocidos por sus siglas en ingles MT *(Machine Translator)*, pertenecen al área del procesamiento del lenguaje natural o lingüística computacional, estas se encargan de desarrollar algoritmos para la traducción de texto o habla de un lenguaje natural a otro. Dentro de este campo se encuentran diferentes enfoques los cuales se pueden observar en la figura 2.

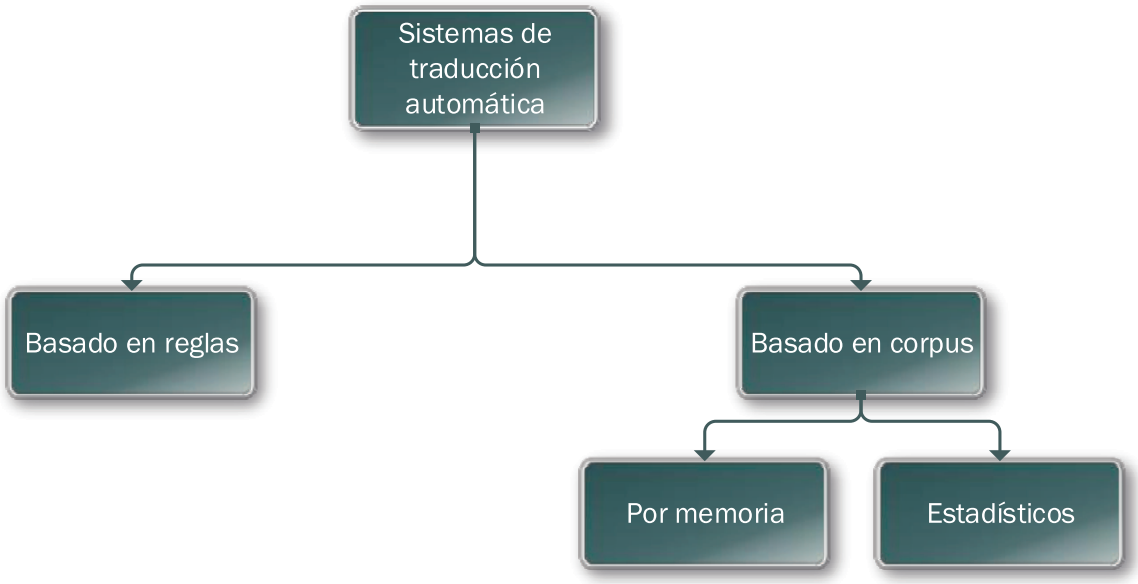


Figura 2: Clasificación de los sistemas de traducción automática

3.4.1. Traducción automática basada en reglas lingüísticas

Los sistemas de traducción automáticos basados en reglas lingüísticas se basan normalmente en el Triángulo de Vauquois que se muestra en la figura 3. En el triángulo se distinguen tres enfoques principales: los enfoques directos, los de transferencia y los de interlengua. Esta pirámide se basa en las diferencias de "longitudes relativas" de los tres componentes de la traducción: análisis, transferencia y síntesis. En la traducción directa que se ubica en la base de la pirámide consiste en realizar una

traducción palabra a palabra desde la lengua origen a la lengua destino. Esta traducción se apoya en diccionarios bilingües. En el intermedio de la pirámide se ubica el enfoque de transferencia, que consiste en realizar un análisis gramatical desde el área sintáctica y algunas veces desde la semántica. Finalmente, en la cima de la pirámide se encuentra la interlengua, que consiste en un análisis semántico profundo.

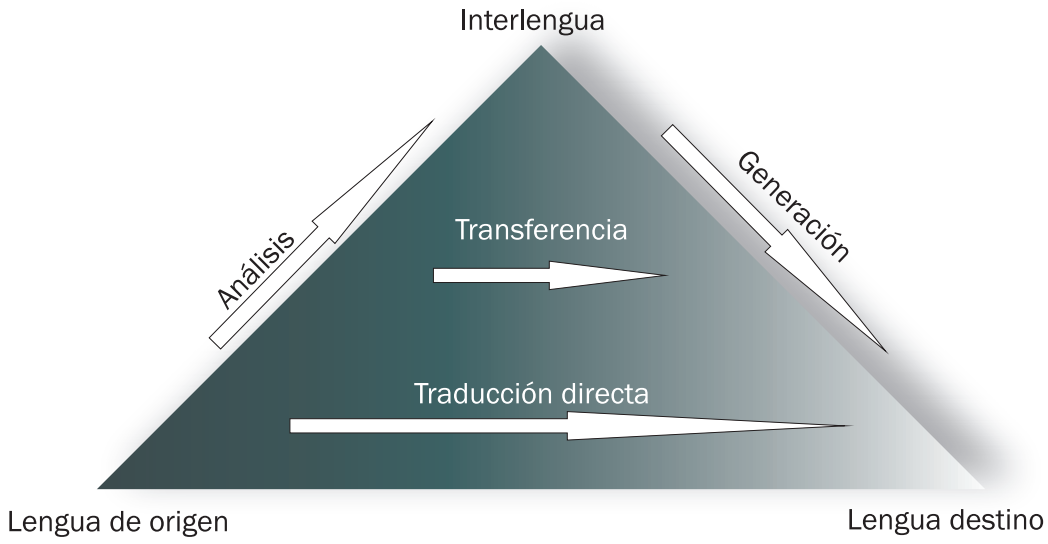


Figura 3: Triángulo de Vauquois

3.4.2. Traducción automática basada en corpus

Los sistemas de traducción basados en corpus, consisten en modelos de traducción que se obtienen a partir del análisis de ejemplos de un corpus paralelo. Este tipo de traducción se divide en dos, por memoria y traducción automática estadística (*Statistical Machine Translation* - SMT). La traducción automática por memoria consiste en una traducción por analogía, es decir, resuelve un problema basándose en la solución de problemas similares existentes en el corpus.

3.4.3. Sistemas estadísticos de traducción

Los primeros sistemas de traducción automática estadística fueron introducidas por Warren Weawer en 1949 [46]. Weawer fue un biólogo y científico de información estadounidense que junto con Claude Shannon consolidaron la teoría de la información. En las décadas de los cincuenta y noventa, el auge de la traducción disminuyó debido a que no se contaba con los equipos para realizar el procesamiento de los datos. Pero al aumentar la capacidad de procesamiento y almacenamiento de las computadoras, los sistemas de traducción automáticos recibieron un nuevo impulso [47]. En 1991 el sistema Candide surgió como un desarrollo que tenia como propósito la traducción entre lenguas utilizando técnicas estocásticas. Este desarrollo se realizó en Nueva York por la empresa IBM y estuvo a cargo de un grupo de investigadores del Thomas J. Watson Center. El sistema fue entrenado con el corpus Hansard de las Actas del Parlamento Canadiense con una cantidad de aproximadamente tres millones de oraciones

en inglés y francés. En el desarrollo del sistema se alinearon oraciones, grupos de palabras y palabras sueltas y después se calcularon las probabilidades de que una palabra de una oración en una lengua correspondiera con otras palabras en la traducción. La mitad de las oraciones traducidas por el sistema eran exactamente como las contenidas en el texto original o tenían el mismo sentido. El sistema no se llegó a comercializar, pero supuso un hito histórico que dio un giro a las investigaciones. Desde 2006, la SMT es la rama de la MT más estudiada. La SMT tiene ventajas con respecto a la MT basada en reglas, tales como, un mejor uso de los recursos, una mayor naturalidad de las traducciones y los sistemas de entrenamiento generados son fácilmente adaptables a otro par de lenguas. El principal inconveniente de la SMT es la dependencia a un corpus paralelo [48].

La fórmulación matemática de un modelo de traducción dentro de la SMT consiste en calcular la probabilidad $p(d|o)$ de que una cadena d de la lengua destino sea la traducción de una cadena o en la lengua origen. Esta probabilidad se calcula aplicando el Teorema de Bayes expresado por la siguiente ecuación:

$$p(d|o) \ \propto \ p(o|d) * p(d)$$

Donde

- $p(o|d)$ es la probabilidad de que la cadena origen sea la traducción de la cadena destino (modelo de traducción)
- $p(d)$ es la probabilidad de ver aquella cadena destino (modelo de lenguaje).

Matemáticamente se calcula la probabilidad de todas las posibles frases y se da como traducción aquella que presenta la probabilidad más alta.

La SMT se puede clasificar de tres maneras: traducción basada en palabras, traducción basada en frases y traducción basada en transductores de estados finitos. Para este trabajo se implementó traducción basada en frases. La SMT basada en frases, resuelve la limitación de la SMT basada en palabras tales como, un mejor uso de los recursos, una mayor naturalidad de las traducciones y los sistemas de entrenamiento generados se adaptan fácilmente a otro par de lenguas. En los sistemas basados en frases, no se consideran de carácter lingüístico, son frases encontradas en el corpus utilizando métodos estadísticos. A estas frases se le suelen llamar comúnmente subfrases. Un proceso de SMT basada en frases o subfrases consta de un modelo de traducción (TM), un modelo de lenguaje (LM) y un decodificador. El modelo de traducción se obtiene a partir del alineamiento entre las frases del corpus paralelo y la extracción de subsecuencias de palabras. El modelo de lenguaje se obtiene por un entrenamiento con la lengua destino. Estos modelos los utiliza un decodificador para generar la traducción. El proceso completo de traducción y evaluación se ejemplifica en la figura 4.

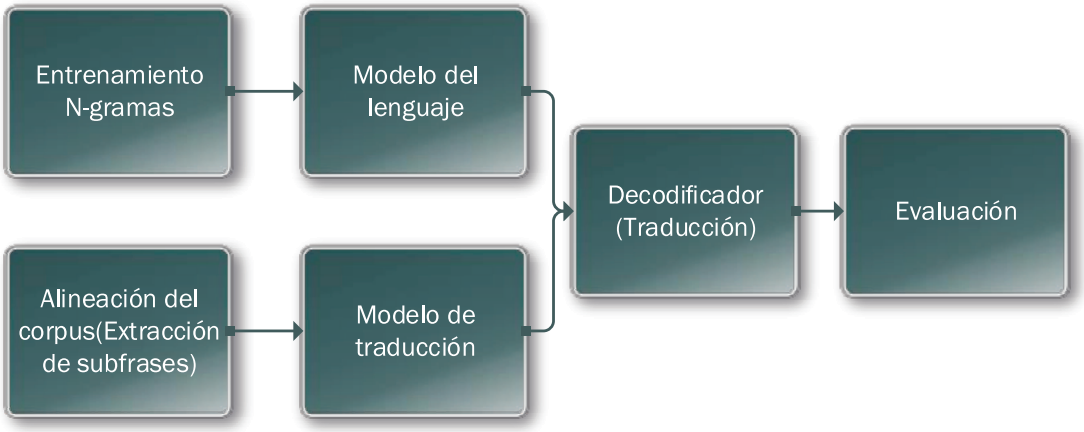


Figura 4: Esquema de traducción basada en frases o subfrases

Para obtener un sistema automático de traducción es necesario primero entrenar modelos del lenguaje tanto del idioma original como el idioma al que va ser traducido. Para obtener los modelos del lenguaje es necesario contar con un *corpus* que contenga un amplio vocabulario del lenguaje con el fin de obtener un modelo matemático robusto, todo este vocabulario se divide en secuencias de una, dos o tres palabras, a estos subconjuntos se les conoce como n-gramas y permitirán el entrenamiento de cada modelo del lenguaje. Para obtener el sistema de traducción automático es necesario contar con un corpus en paralelo tanto de la lengua original como su traducción, para conseguir un buen modelo de traducción se debe realizar el alineamiento de estos dos corpus y con esta información se obtiene tanto el modelo del lenguaje como el sistema de traducción automática. Una vez obtenido el LM y el MT el sistema, se calculan las probabilidades para todas las posibles secuencias, con esta información se realice la decodificación determinar cuál traducción es la que presenta mayor probabilidad [49].

Modelo del lenguaje

Uno de los modelos de lenguaje probabilísticos más utilizado para la SMT, es el modelo basado en n-gramas (subdivisiones del corpus en un tamaño definido). Estos modelos calculan la probabilidad de una palabra basada en las palabras anteriores, es decir, se basa en la muestras pasadas para predecir la palabra siguiente. El grado del modelo (n), indica que el modelo se basa en el contexto de las $n-1$ palabras anteriores. Si el modelo es de segundo orden ($n = 2$) se denomina bigrama, si es de tercer orden ($n = 3$) se denomina trigrama. Para hallar la probabilidad, se usa la regla de la cadena expresada en la ecuación 1.

Utilizando como ejemplo una frase compuesta por 3 palabras en total.

$$p(a, b, c) = p(a) * p(b|a) * p(c|b, a) \tag{1}$$

Donde:

- $p(a)$ Probabilidad de que palabra a se encuentre dentro de la frase.
- $p(b|a)$ Probabilidad de que palabra b se encuentre dentro de la frase dada la palabra a.

- $p(c|b,a)$ Probabilidad de que palabra c se encuentre dentro de la frase dada las palabra a y b.

Cada uno de los factores de la ecuación 1 se calcula por medio de la ecuación 2

$$p(b|a) = \frac{frecuencia\,de\,las\,subfrases\,(a,b)}{frecuencia\,de\,las\,subfrases\,(a,x)} \tag{2}$$

$$p(c|b,a) = \frac{frecuencia\,de\,las\,subfrases\,(a,b,c)}{frecuencia\,de\,las\,subfrases\,(a,b,x)}$$

Pero al aplicar la regla de la cadena a textos muy amplios, el conteo de las frecuencias es demasiado extenso, así que se recurre a aplicar la Cadena de Márkov. La Cadena de Márkov consiste en una serie de eventos, en la cual la probabilidad de que ocurra un evento depende del evento inmediato anterior, este postulado se expresa en la ecuación 3

$$p(\omega_1,\omega_2,\omega_3,....,\omega_n) \approx \prod_{i=1}^{n} p(\omega_i|\omega_{i-k},.....,\omega_{i-1}) \tag{3}$$

Donde:

- n es el número total de palabras.
- i es un contador que itera por cada palabra de la frase.
- k es el número total de gramas ($k-gramas$).

3.5. Métricas para evaluar los sistemas de traducción automática

Con el fin de obtener una medida cuantitativa de que tan bien opera un sistema de traducción automático se han creado diferentes métricas. Estas se basan en los n-gramas, los cuales son secuencias de palabras agrupadas o no, que permiten identificar una o varias frases con el fin de realizar una traducción.

Para esto se cuenta con 1-grama que es una agrupación de palabras, los 2-grama o llamados bigramas que son frases compuestas por dos palabras y los trigramas o 3-grama, que son frases que contienen 3 palabras. A continuación se presentan las métricas empleadas en este trabajo.

3.5.1. BLEU

BiLingual Evaluation Understudy (BLEU) es una de las métricas de evaluación usada para evaluar sistemas de traducción automática. Es una métrica que presenta un costo computacional bajo, además de ser independiente del lenguaje y tiene una gran correlación con las evaluaciones manuales [50]. BLEU compara los n-gramas de la frase generada por el sistema de traducción con los n-gramas de la frase de referencia, contando el número de n-gramas que coinciden independientemente de la posición. La BLEU se puede calcular utilizando más de una traducción de referencia, lo cual permite una mayor robustez a la medida frente a traducciones libres realizadas por humanos. La BLEU se calcula mediante la siguiente expresión

$$BP = \left\{ \begin{array}{c} 1 \; si \; c > r \\ e^{(1-r/c)} \; si \; c \leq r \end{array} \right\}$$

$$BLEU = BP.exp\left(\sum_{n=1}^{N} w_n log(p_n)\right)$$

Donde :

- N: Orden de los n-gramas calculados.
- BP: Es un factor que penaliza las traducciones que sean más cortas que su frase original.
- Pn: Precisión.
- Wn: Peso uniforme de la forma $Wn = \frac{1}{N}$.

3.5.2. WER

Porcentaje de Palabras Erróneas o Word Error Rate (WER) [51]. Está métrica mide el porcentaje mínimo de palabras que se insertaron, eliminaron o sustituyeron al realizar la traducción respecto a la frase de referencia. El WER se calcula mediante el conteo del número de inserciones (palabras introducidas en la salida de texto por el sistema de reconocimiento), borrados (palabras perdidas por el sistema de reconocimiento) y sustituciones (falta de reconocimiento de una palabra por otra) de palabras cuando se compara la traducción con la referencia. Esta medida se basa en la distancia de edición o de Levenshtein. La expresión para calcular el WER en cada frase de salida del traductor con respecto a la frase de referencia es:

$$WER = \frac{S + D + I}{S + D + C}$$

Donde:

- D Número de palabras borradas.
- S Número de palabras sustituidas.
- C Número de palabras correctas.
- I Número de palabras insertadas.

También se suele emplear el Word Accuracy (W_{Acc}), calculado como :

$$W_{Acc} = 1 - WER$$

4. DESARROLLO

En este capítulo se mostrarán los desarrollos realizados, partiendo de la sección 4.2, donde se muestran la creación de la base de datos, realizando una descripción completa sobre cómo se capturo la misma y cómo se organiza la información, para luego pasar por la sección 4.3 donde se muestran los descriptores utilizados para realizar el reconocimiento de las señas y respectiva clasificación utilizando algoritmos de aprendizaje de máquina, se explica cómo se realiza el entrenamiento y validación. Después se muestra el sistema de traducción implementado en la sección 4.4 con sus respectivas modificaciones en las variables de sintonización al igual que las pruebas que se realizaron y las métricas utilizadas para verificar su funcionamiento.

4.1. Desarrollo metodológico

La metodología planteada para el desarrollo de este trabajo se muestra en la figura 5.

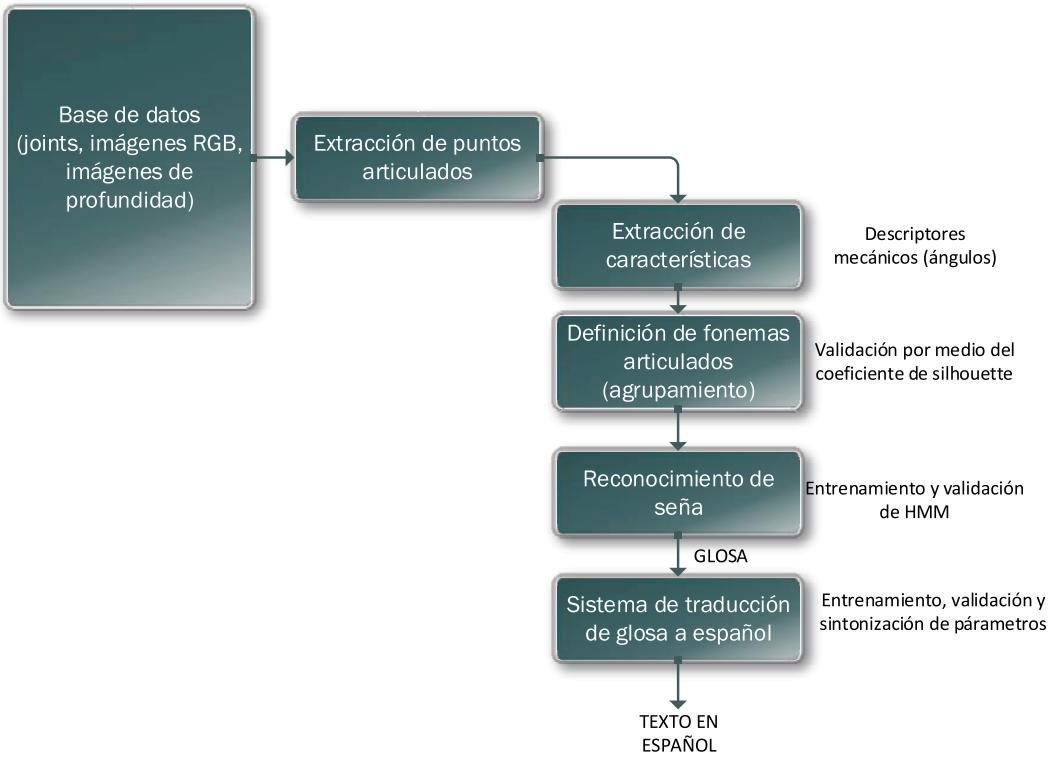


Figura 5: Metodología planteada para la realización de este trabajo

Con el fin de cumplir con el objetivo general se planteó una metodología estructurada que permitió concluir con la investigación.

Para el primer objetivo específico el cual fue: crear una base de datos anotada compuesta por imágenes y videos de intérpretes de lenguaje de señas colombiano en un dominio de aplicación específico, se realizó la captura por medio de una cámara de profundidad, con esta información se continuo con el segundo objetivo específico el cual es determinar un espacio vectorial de representación para la ex-

tracción de características a partir de los puntos articulados entregados por el sensor de profundidad. Para esto se realizó una búsqueda en el estado del arte con el fin de determinar la técnica que permitió extraer las características relevantes y diferenciables para las diferentes señas, adicional se propusieron nuevos métodos de extracción de características con el fin de determinar el que presenta un mejor comportamiento. Los métodos planteados fueron descriptores mecánicos, descriptores utilizados para determinar las posiciones y orientaciones en un objeto de interés dentro de una imagen, además de que permitieron eliminar la dependencia inter-subjetiva de las señas respecto a las medidas corporales de cada persona. Para alcanzar el tercer objetivo específico. fórmular una metodología para la creación de un diccionario de "fonemas" articulados utilizando técnicas de agrupamiento no supervisado. Se utilizó k-means y el coeficiente de silhouette con el fin de determinar la cantidad mínima de movimientos que representan el universo de señas con las que se cuenta en la base de datos. Los fonemas articulados calculados se utilizan para cumplir con el cuarto objetivo específico, el cual es, determinar una metodología para el reconocimiento automático de la lengua de señas utilizando técnicas de aprendizaje de máquina que hagan uso del diccionario de fonemas articulados que se creó en el anterior objetivo. En este punto se utilizó modelos ocultos de Markov con el fin de determinar la seña aislada o el conjunto de señas que realizó el intérprete. Una vez identificada la seña se pasó la seña clasificada por un sistema de traducción y se midió la calidad de traducción por medio de BLEU y WER para así cumplir con el quinto objetivo específico el cual era validar estadísticamente el rendimiento de las metodologías planteadas mediante métricas empleadas en el campo del procesamiento del lenguaje natural.

4.2. Creación de la base de datos

La creación de la base de datos es uno de los puntos fundamentales dentro del proyecto ya que actualmente no existen bases de datos capturadas por medio del Kinect One del lenguaje de señas colombiano. Ésta base de datos se utilizó para la extracción, clasificación y reconocimiento de la seña. Los pasos que se siguieron para la creación de la base de datos fueron los siguientes:

4.2.1. Definición de niveles de dificultad

Para la creación de la base de datos fue necesario reunirse con dos intérpretes del lenguaje de señas y una persona experta en la temática. Con la ayuda de ellos se pudo observar que todas las señas presentan movimiento, solo que algunas llevan las manos a alguna parte del cuerpo, permanecen allí y luego finalizan, en otras señas, las manos van a alguna parte del cuerpo y generan un movimiento adicional antes de finalizar. También se aprecia que dentro de este lenguaje de señas existen algunas que presentan oclusiones parciales o totales entre manos. Además, solo es necesario capturar el tronco superior ya que las señas no dependen de las piernas. También se apreció que todas las señas empiezan desde «silencio» que corresponde a dejar los brazos abajo en lengua de señas como se puede observar en la figura 6.

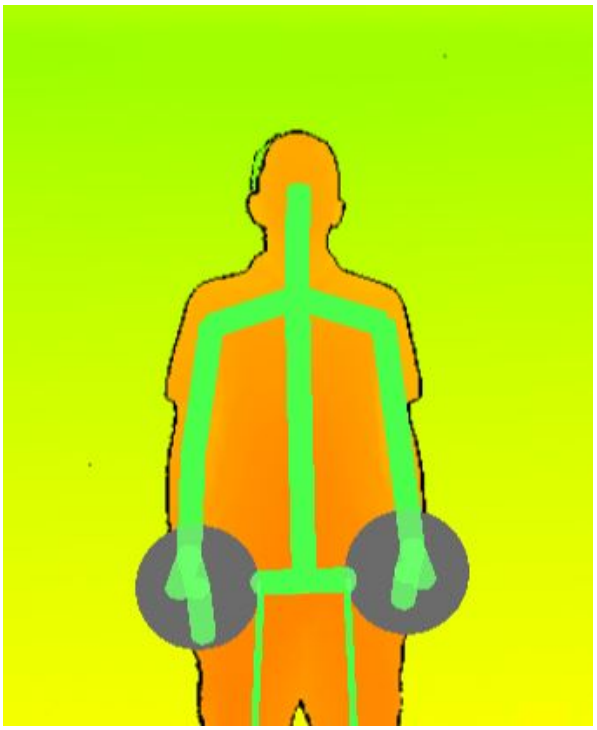


Figura 6: Imagen de profundidad del «silencio» en lenguaje de señas

Teniendo todo lo anterior presente se clasificaron las señas en cuatro niveles de dificultad básico, fácil, medio y difícil. La forma de determinar cada uno de los niveles definidos se explicará a continuación:

Nivel básico: Las señas del nivel básico se caracterizan por no tener oclusiones entre las manos, no presentan movimiento entre los dedos mientras se realiza el movimiento, llegan las manos a una posición indicada y permanecen allí sin realizar ningún movimiento adicional hasta concluir la seña. Una de las señas de nivel básico se observa en la figura 7.

Figura 7: Ejemplo de seña clasificada en el nivel básico

Nivel fácil: En el nivel fácil se presentan señas que no tienen oclusión entre manos, los dedos permanecen quietos durante la realización de la seña, pero a diferencia del nivel anterior una vez

realizada la seña la mano se mueve haciendo un gesto adicional para luego llegar a la posición final. Un ejemplo de este nivel se pueden observar en la figura 8.

Figura 8: Ejemplo de seña clasificada en nivel fácil

Nivel medio: El nivel medio se caracteriza por presentar oclusiones parciales entre las manos, existe movimiento en los dedos mientras se realiza otro movimiento, adicional se llega a un punto del cuerpo donde se realiza un movimiento repetitivo, los dedos no cambian de posición mientras se realiza esta acción. Un ejemplo de las señas del nivel medio se puede observar en la figura 9.

Figura 9: Ejemplo de seña clasificada en nivel medio

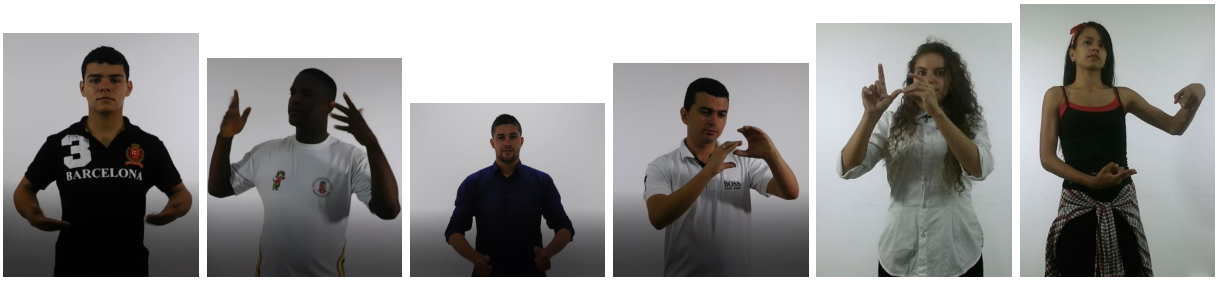
Nivel difícil: En el nivel difícil se clasificaron las señas que presentan oclusiones parciales entre manos, los dedos cambian de posición mientras se realiza la seña y cuenta con la particularidad de llegar a una parte del cuerpo y realizar una acción adicional mientras los dedos cambian de posición. En la figura 10 se pueden apreciar un ejemplo de una señas clasificada como difícil.

Figura 10: Ejemplo de seña clasificada en nivel difícil

4.2.2. Captura de la base de datos

Con los niveles de dificultad definidos se prosiguió a realizar la captura de 199 señas aisladas, las señas capturada son glosas que representan preguntas típicas de un centro de información de una entidad de educación superior, tales como ubicaciones de lugares, costos de matricula, documentos necesarios para realizar tramites, medios de transporte, entre otros. Adicional a las señas aisladas se capturaron 51 señas compuestas[2] que presentan el mismo contexto de las señas. La captura se realizó por medio del Kinect One y una cámara digital marca Nikon a una resolución de 1280x720 píxeles, esto con el fin de obtener información de otro sensor y dejar información para futuros proyectos.

La base de datos capturada consta de seis actores, dos mujeres y cuatro hombres los cuales presentan diferentes estaturas esto con el fin de obtener variabilidad en los datos. De los actores 5 son nativos y uno es un intérprete. Los actores se pueden apreciar en la figura 11.



(a) Actor 1 (b) Actor 2 (c) Actor3 (d) Actor 4 (e) Actor 5 (f) Actor 6

Figura 11: Actores de la base de datos

La base de datos capturada queda almacenada en archivos con extensión XEF, extensión propia del SDK(Software Development Kit) de Kinect que permite almacenar toda la información (puntos articulados del cuerpo, la imagen de la cámara de color y la imagen de profundidad) capturada por el sensor en un solo archivo. Un ejemplo de la información capturada y almacenada se puede observar en la figura 12.

[2]señas compuestas: son aquellas que representan una frase completa en el español

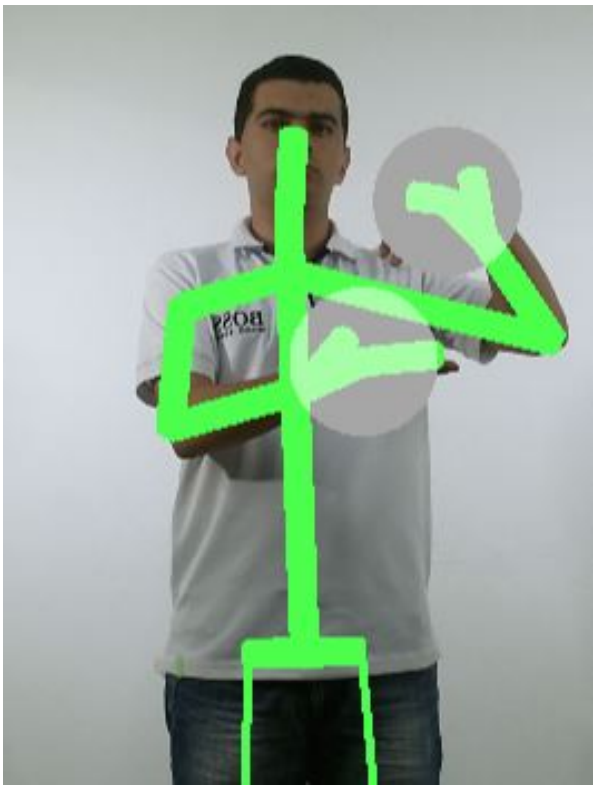


Figura 12: Imagen capturada de la base de datos

4.2.3. Extracción de información de archivos XEF

Debido a que los archivos capturados con el Kinect se encuentran comprimidos, es necesario extraer la información de los puntos articulados, imágenes de profundidad y de color para poder trabajar los algoritmos planteados. Por lo tanto, se utiliza el SDK del kinect y un *toolbox* utilizado en [52]. Esta herramienta permite extraer la información *frame* por *frame* de un archivo XEF y guardar la información en archivos con extensión txt para los puntos articulados del cuerpo y bmp para las imágenes, el modelo de extracción se puede observar en la figura 13

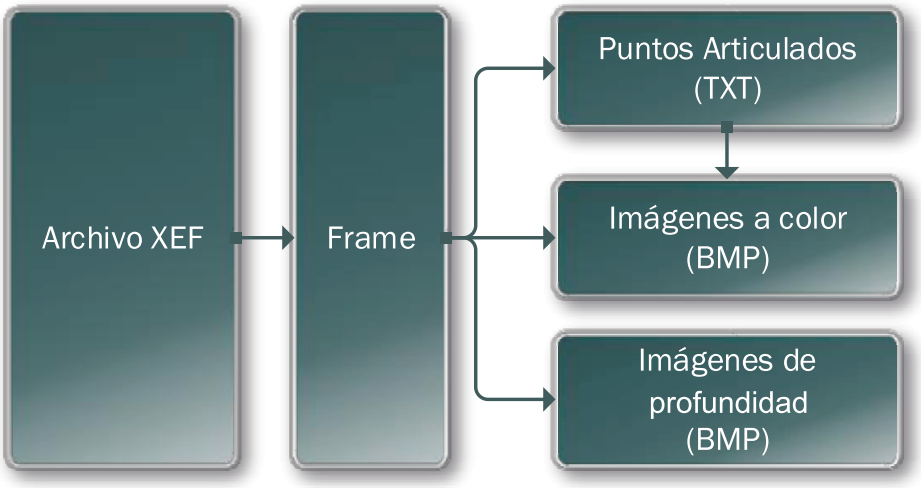


Figura 13: Extracción de información de archivos XEF

Basados en los códigos de SDK se les realiza una modificación con el fin de cambiar de coordenadas en pixeles a coordenadas en metros, para así contar con toda la información y poder aplicar diferentes algoritmos de extracción de características.

4.2.4. Errores en la captura de datos

Al realizar la extracción de los datos se observó que algunos datos presentan errores ya que aparecen desincronizados entre los datos la cámara de profundidad, la cámara a color y los puntos articulados. Por lo tanto, fue necesario realizar el proceso de extracción de una manera exhaustiva. Por otro lado, como se puede observar en la figura 14, los puntos articulados en algunas partes de la seña no corresponden a la seña realizada por la persona.

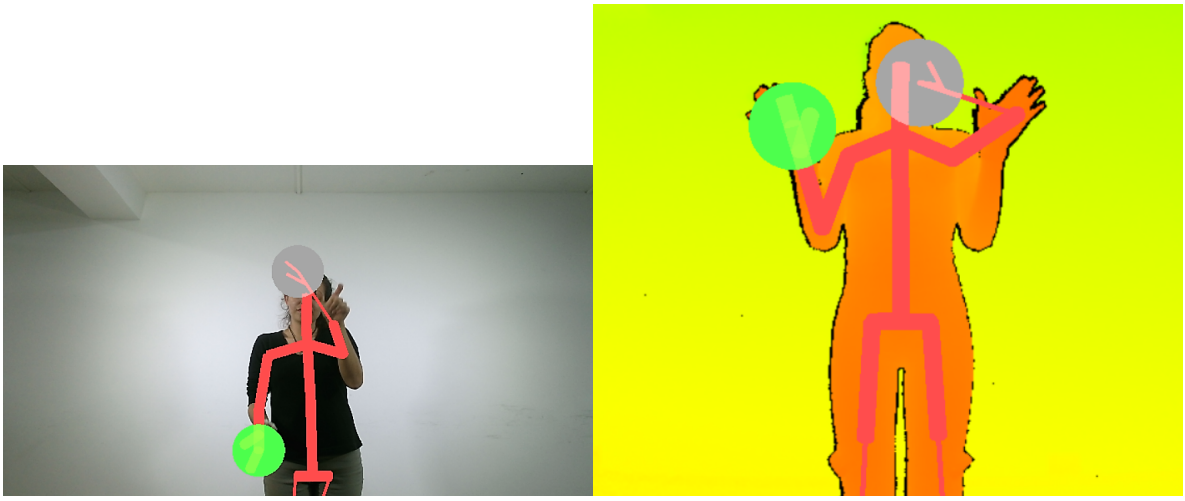


Figura 14: Errores que presenta el sistema de captura

4.3. Reconocimiento de señas

Con el fin de comprobar la metodología planteada se toman 30 señas aisladas y 10 señas compuestas. Con esta subdivisión se plantean tres descriptores mecánicos que sean invariantes a la escala (que sin importar el tamaño de la persona el comportamiento tienda a ser el mismo). Con la información obtenida por medio de los descriptores se procede a definir los mínimos movimiento que conforman las señas (fonemas articulados), para esto se utilizó K-means y se hace variar la cantidad de grupos de 1 a 100, calculando en cada uno de los casos el coeficiente de silhouette. Se decide variar de 1 a 100 la cantidad de grupos ya que lo que se busca en este trabajo es definir los fonemas articulados del lenguaje de señas colombiano y se tiene como base los lenguajes hablados que presentan entre 20 a 60 fonemas. Una vez obtenido el conjunto de grupos por los que pasaron cada una de las señas, se realiza el entrenamiento y validación de un conjunto de modelos ocultos de Márkov, para esto se utiliza una validación cruzada en seis partes. Para cada una de las frases se toman cinco actores y se valida con uno, este proceso se repite seis veces. La metodología para el reconocimiento de la seña realizada se presenta en la figura 15. En ella se puede apreciar que dentro de los descriptores utilizados se plantean 3 descriptores cada uno de ellos lo que busca es quitar la dependencia de la seña respecto a las medidas antropométricas de la persona que realice la seña, para intentar lograr esto se busca primero obtener un mismo punto de referencia y trabajar con el cálculo de ángulos tanto en coordenadas cartesianas como el cálculo de coordenadas esféricas.

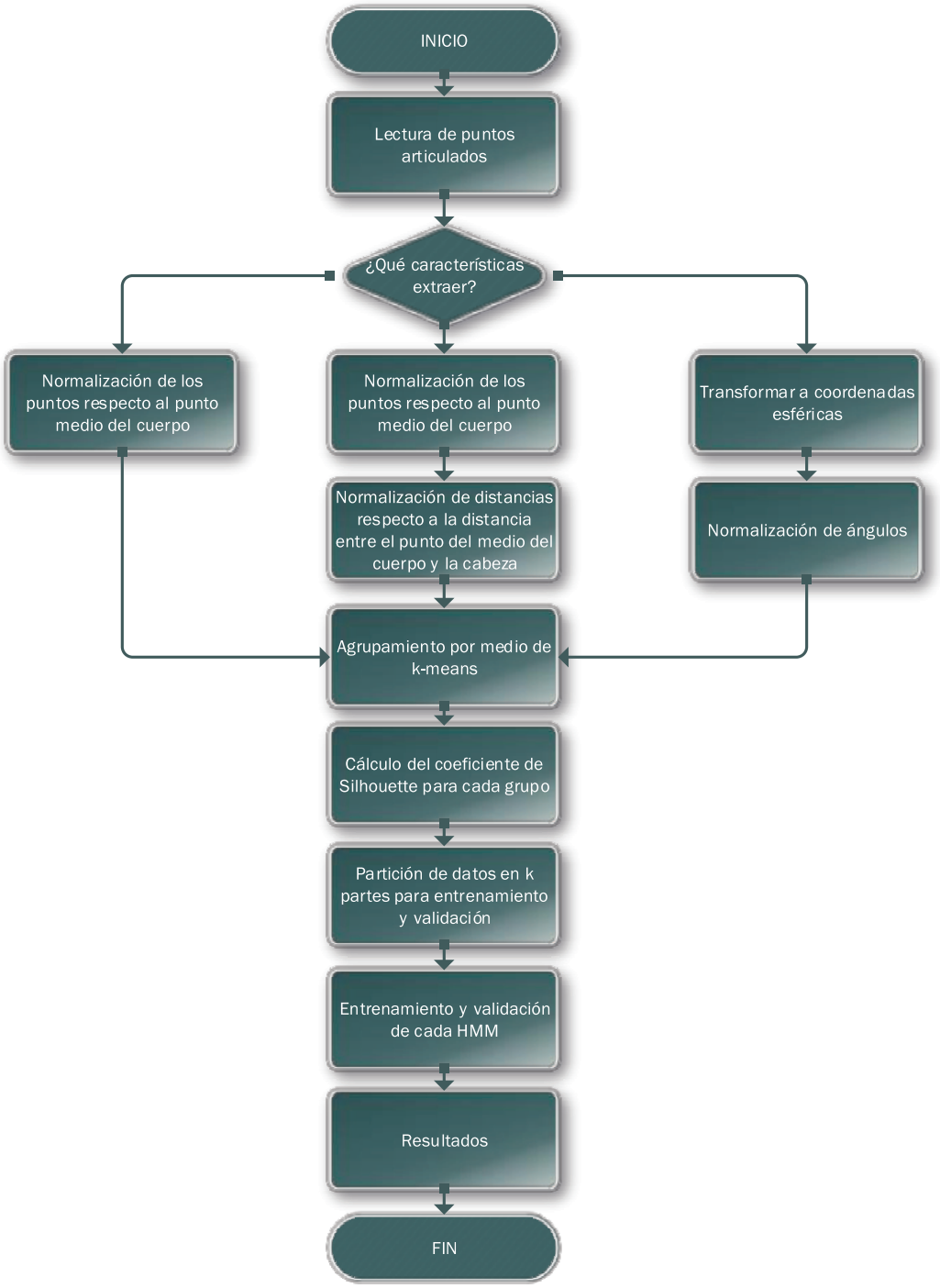


Figura 15: Metodología planteada para el reconocimiento de señas

4.3.1. Extracción de características

Con el fin de realizar la extracción de característica se propone utilizar la información capturada por el Kinect, para esto se realizó una modificación sobre los códigos del SDK del Kinect, permitiendo

así que se ingrese un archivo en formato XEF y se pueda extraer las imagen RGB, la imagen de profundidad y los puntos articulados en coordenadas cartesianas.

El Kinect One entrega 25 puntos articulados como se puede observar en la figura 16, de los cuales solo se seleccionaron 16 de esto ya que el tronco inferior del cuerpo no agrega información, también cabe resaltar que el «silencio» en el lenguaje de señas equivale a colocar las manos abajo.

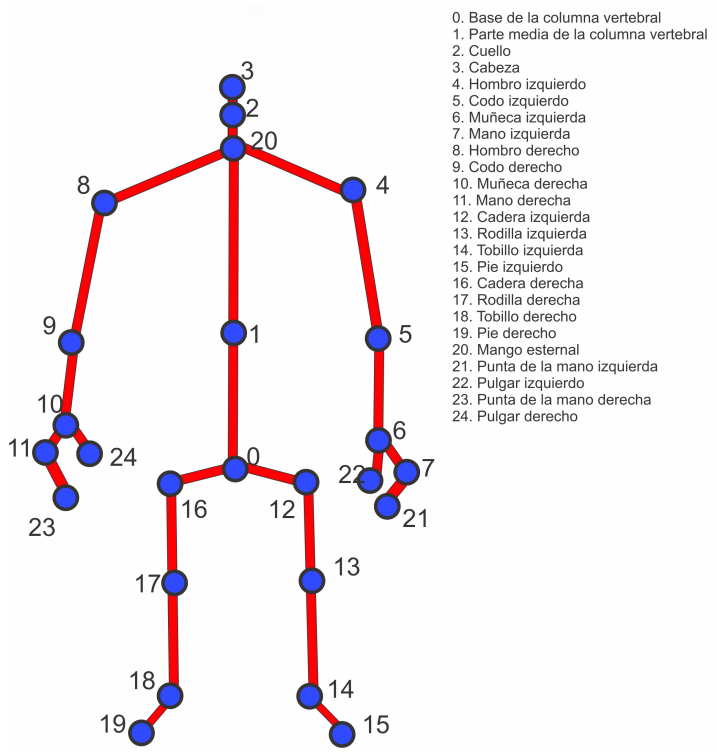


Figura 16: Puntos articulados que entrega el kinect One

Los puntos articulados que entrega el Kinect presentan dos grandes problemas. El primero es que dependiendo de donde se ubique la personas las coordenadas van a cambiar, ya que el sensor presenta un origen de coordenadas que no depende de la ubicación de la persona. El segundo es que cuando se realiza la captura de la seña las personas seleccionadas para crear la base de datos presenta diferentes medidas antropométricas. Por tanto es necesario crear algoritmos que eliminen la dependencia de los datos respecto a la posición en la que se encuentra el actor y que también permitan eliminar la dependencia de las medidas del cuerpo de quien realice la seña. Ahora se explicarán los descriptores utilizados para eliminar estas dependencias.

Cambio de punto de origen El origen de coordenadas que presenta el Kinect se ubica en la mitad de la imagen, por lo tanto, si una persona se ubica cerca a este o lejos y realiza una seña según el punto de origen del sensor Kinect, la seña es diferente. Por lo tanto se toma un punto equidistante del cuerpo como nuevo origen del sistema. El punto seleccionado para el nuevo origen de coordenadas es el punto que el Kinect llama como punto medio de la columna vertebral como se observa en la figura 17.

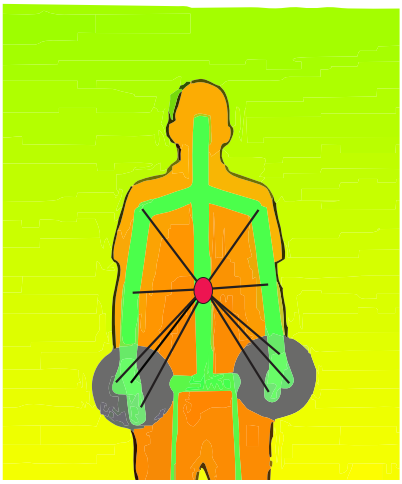


Figura 17: Nuevo origen de coordenadas

Para cambiar el origen de coordenadas se utilizó la siguiente fórmula:

$$Pn(i) = P(i) - Pn$$

Donde

- Pn es el nuevo punto con el cambio de origen de coordenadas.
- $P(i)$ es el punto con el origen de coordenadas original.
- Pn es el punto medio de la columna vertebral.

Con estos vectores en coordenadas del mundo se realiza los primeros agrupamientos, en la figura 18 se observa el movimiento en el eje z del punto articulado de la mano derecha para los seis actores realizando la misma seña. Se puede apreciar que los puntos articulados cuentan con una dinámica parecida entre los puntos articulados eliminando la posición en que se ubica la persona dentro de la escena, pero también se observa que el movimiento realizado presenta diferencia en magnitud, por lo tanto es necesario realizar una normalización de las magnitudes y así poder eliminar la dependencia de los datos respecto a las medidas que presente cada uno de los actores que realizaron la seña.

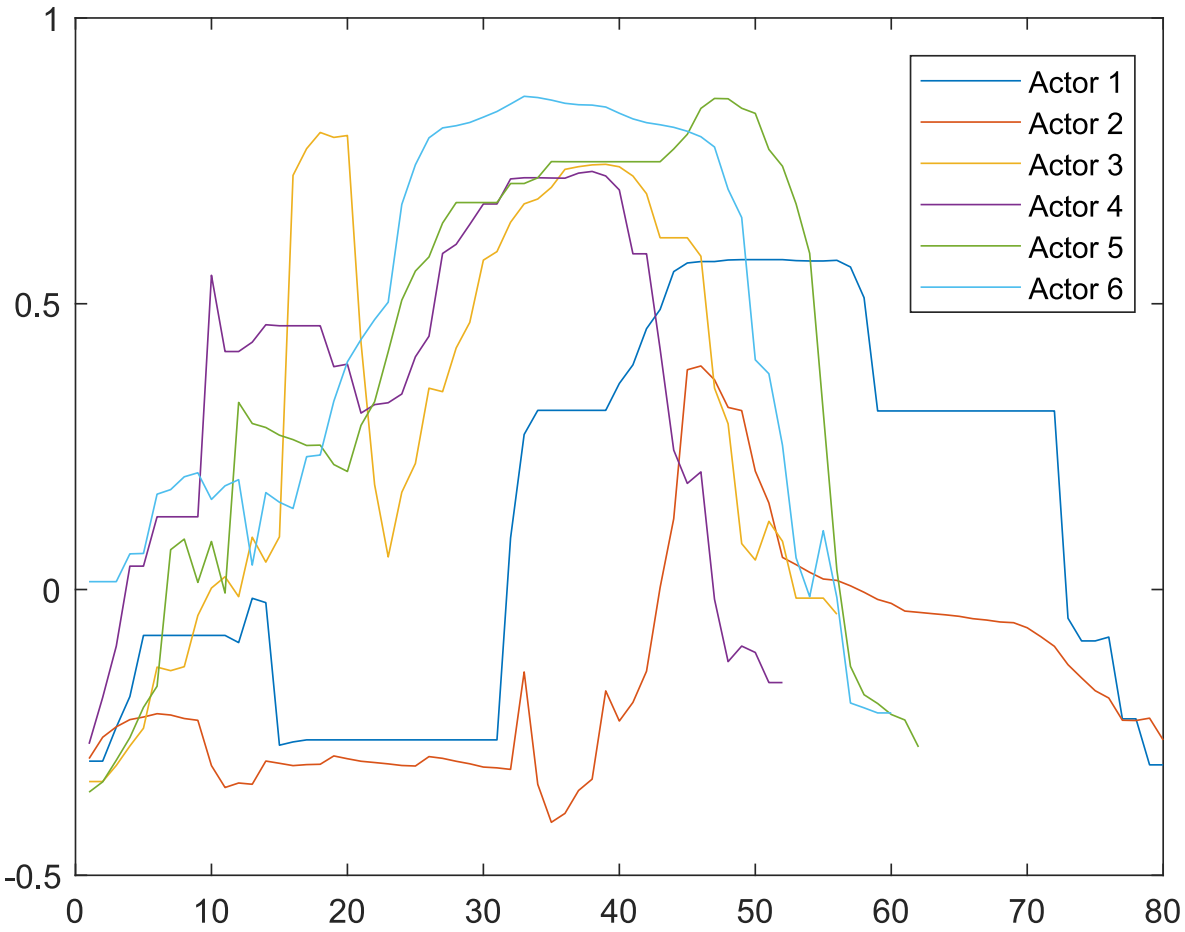


Figura 18: Movimiento en el eje Z del punto articulado de la mano derecha realizado por seis actores

Cambio de punto de origen y normalización de la magnitud Una vez definido un punto de origen es necesario buscar una forma de normalizar las medidas antropométricas. Para esta labor se dividen cada uno de los datos respecto a la distancia entre el punto medio del cuerpo y la cabeza como se observa en la figura 19.

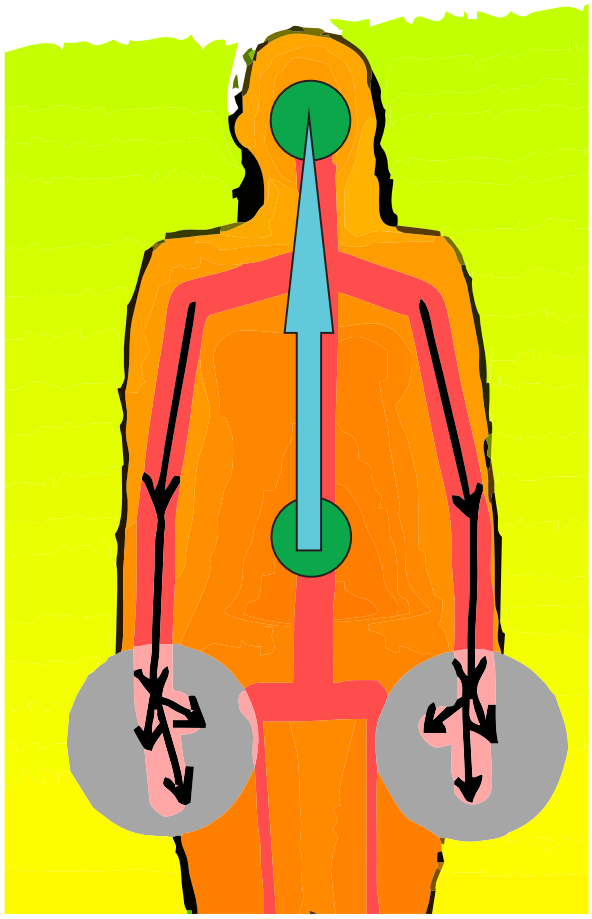


Figura 19: Normalización de datos respecto a puntos claves

La fórmula que se utilizó para realizar este proceso es la siguiente:

$$PN(i) = \frac{P(i) - Pn}{||Pc - Pn||}$$

Donde:

- $PN(i)$ es el nuevo punto normalizado y con cambio de origen.
- Pc es el punto de la cabeza.
- $P(i)$ es el punto con el origen de coordenadas original.
- Pn es el punto medio de la columna vertebral.

En la figura 20 se muestra el movimiento en el eje z de la mano izquierda para los seis actores. Se observa la misma característica pero ya normalizada respecto al punto central del cuerpo, en esta se puede apreciar que la dinámica sigue siendo parecida y que una vez normalizado se pasa de una

diferencia máxima entre los actores de 0.75 a 0.16. Para el cálculo de esta diferencia se utilizo *dynamic time warping* entre todos los actores.

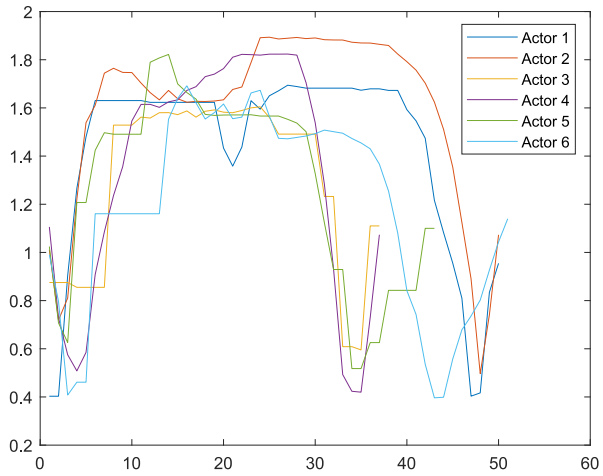


Figura 20: Datos con cambio de referencia sin normalizar

Ángulos normalizados y distancias Otra forma de eliminar la dependencia de las coordenadas respecto a su magnitud y ubicación original es cambiar de sistema de coordenadas o trabajar con ángulos relativos entre los diferentes vectores del cuerpo, por lo tanto se calculan los 10 vectores que permiten describir el movimiento que presenta los dedos y las manos como se explicó en la sección 3.1, todo esto con el fin de determinar la dinámica de este tipo de lenguaje. Los vectores calculados se observan en la figura 21.

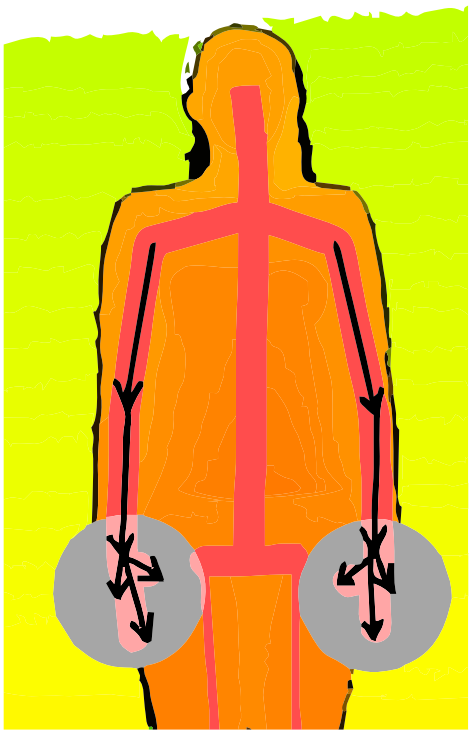


Figura 21: Vectores calculados para realizar el entrenamiento

Con estos vectores se calculan ocho ángulos, los cuales se forman entre los diferentes vectores como se pueden apreciar en la figura 22.

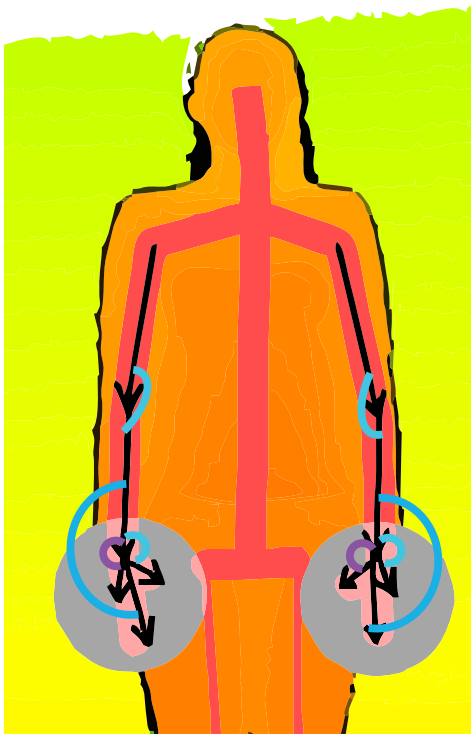


Figura 22: Ángulos calculados con los vectores de las manos y brazos

Estos ángulos son calculados con la siguiente fórmula

$$\theta_j = arccos\left(\frac{v_{j,i}(t) * v_{j,k}(t)}{|v_{j,i}(t)| * |v_{j,k}(t)|}\right)$$

Donde:

- θ_j ángulo entre los vectores de la figura 22
- $v_{j,i}(t)$ vector que une los puntos j e i
- $v_{j,k}(t)$ vector que une los puntos j e k
- $|v_{j,i}|(t)$ norma del vector que une los puntos j e i
- $|v_{j,k}|(t)$ norma del vector que une los puntos j e k

Esta información se le extrae a cada *frame* generando un vector de características de 17000xG para las señas aisladas de 4500xG (donde G cambia según el descriptor utilizado), cabe resaltar que cada seña presenta diferentes duraciones de tiempo, por lo tanto el vector de características tiene una duración diferente para cada una de las señas. En cada seña se le elimina el silencio, ya que esta información no es relevante pero si ocupa espacio a la hora de analizar los datos.

4.3.2. Fonemas articulados

Se definió como fonema articulado al conjunto mínimo de movimientos que componen una seña, este nombre surge haciendo una comparación con los lenguajes hablados y la forma en que se analizan desde el punto del procesamiento del lenguaje natural, ya que toda palabra en los lenguajes hablados está compuesta por fonemas. Los fonemas son las unidades sonoras mínimas que conforman toda una palabra.

Los fonemas facilitan la identificación de las palabras al igual que las señas. Por ejemplo, es más fácil contar con 22 fonemas y conocer la probabilidad en que se presenta cada uno y no crear un clasificador, que tenga 88000 palabras. Por lo tanto es más sencillo identificar los mínimos movimientos (fonemas articulados) que conforman todas las señas, agruparlas y luego reconocer el orden en que se realizan. Y no identificar cada uno de los signos que presenta el lenguaje de señas.

La idea principal de los fonemas articulados sale del concepto de los lenguajes hablados, los cuales presentan sonidos mínimos que conforman una palabra, estos sonidos mínimos son llamados fonemas. Con el fin de obtener los micromovimientos, o fonemas articulados, se toman todas las señas y se realiza un agrupamiento con cada una de las características extraídas utilizando el algoritmo de K-means y se toma como punto de partida de la búsqueda de los movimientos mínimos la cantidad de fonemas que presenta una lengua hablada, que puede variar entre 20 a 60. Por lo tanto, se realiza diferentes agrupamientos variando la cantidad de grupos entre 1 a 100. A cada uno de estos agrupamientos se les calcula la cohesión entre datos y la separabilidad entre los grupos utilizando el coeficiente de silhouette, explicado en la sección 3.2.2, con el fin de determinar un punto de partida para buscar los fonemas articulados del lenguaje de señas colombiano y así poder determinar las matrices articulatorias y segmental que se mencionaron en la sección 3.1.

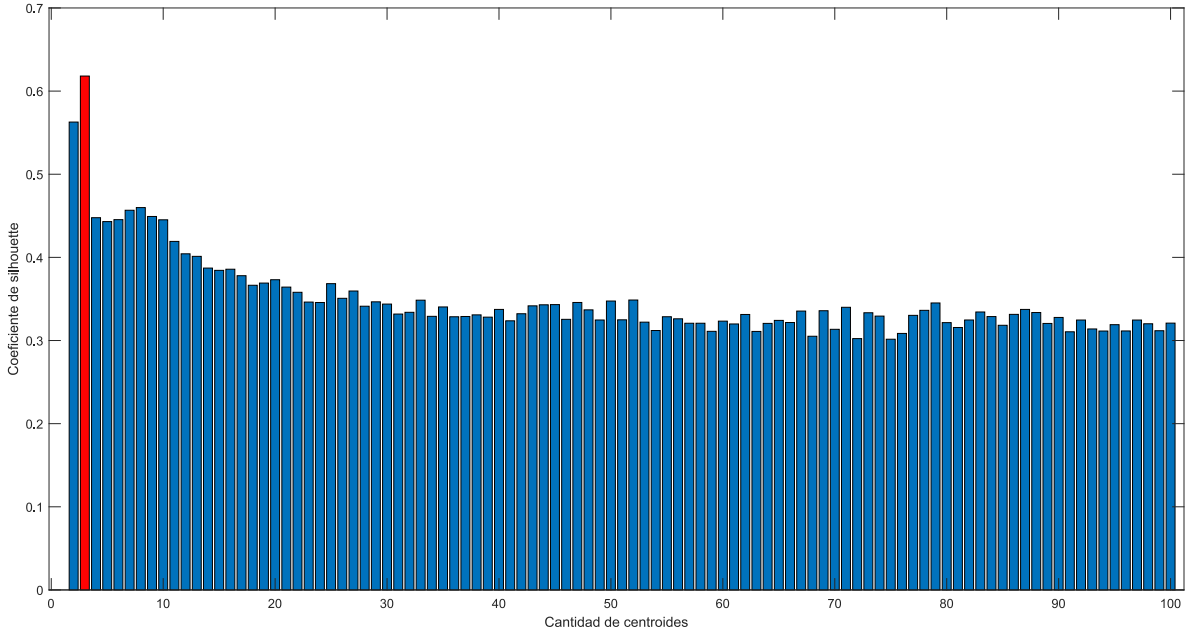


Figura 23: Evaluación de los agrupamientos con el coeficiente de silhouette

En la figura 23 se puede apreciar la evaluación del coeficiente de silhouette para los 100 grupos. Lo que se busca por medio de este coeficiente es realizar una búsqueda que permita determinar la cantidad mínima de movimientos necesarios para representar cualquiera de las señas con las que se cuenta en la base de datos, para así cumplir con el objetivo de determinar el diccionario de fonemas articulados. Por medio de este coeficiente se determiná la cohesión (que tan cercanos están los datos de un mismo grupo) y la separación (distancia entres los agrupamientos), entre mayor sea el coeficiente, mayor sera la cohesión entre los datos y la separación entre los grupos.

4.3.3. Algoritmo de reconocimiento

Para realizar el reconocimiento de las señas en este trabajo se utilizá HMM. Como ya se ha mencionado, esta técnica permite modelar la dinámica propia de la lengua de señas y de acuerdo a [13, 48, 14, 16] ha demostrado ser efectiva al momento de reconocer lenguas de señas en distintos idiomas.

El entrenamiento de las HMM se realiza por medio de las observaciones entregadas por los diferentes agrupamientos realizados por medio de K-means, para el entrenamiento y validación se realiza una validación cruzada de 6 partes, entrenando cada HMM con 5 actores y validando uno, y se realizan cambios en la cantidad de observaciones del modelo de Márkov desde 1 hasta 100 con el fin de validar si

el coeficiente de silhoutte entregado en el punto anterior presenta la mejor clasificación a nivel de señas, y se definen cada HMM con 2 estados con una distribución uniforme. Para validar el reconocimiento se toma el nuevo dato y se pasa por cada HMM previamente entrenada y se determina cual fue el modelo que entrega la mayor verosimilitud, este modelo determina a que clase pertenece, el esquema de validación se puede apreciar en la figura 24.

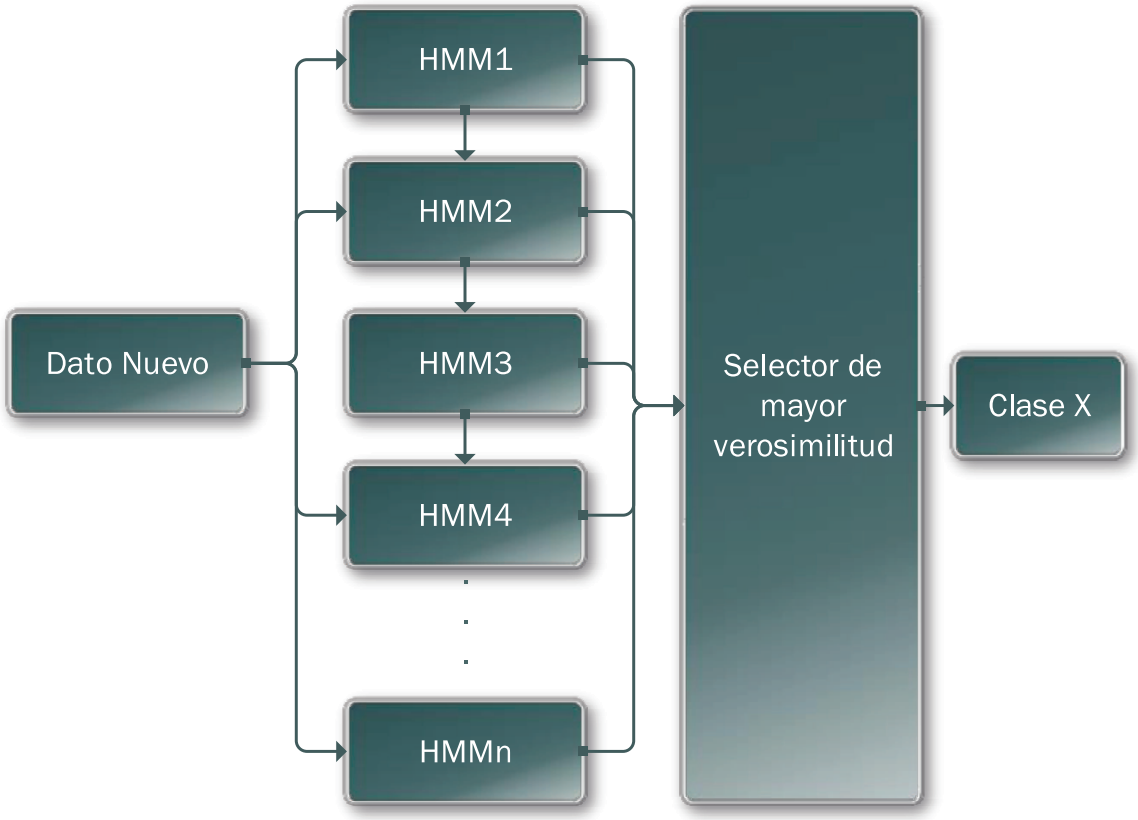


Figura 24: Esquema de validación utilizando HMM

La modelo matemático para esta validación es el siguiente:

$$clase = max(P(HMM_i(x))$$

Donde

- x es el conjunto de observaciones evaluadas.
- HMM_i es cada una de las HMM entrenadas para cada una de las señas.
- P la probabilidad de que el conjunto de observaciones x al ser evaluadas en cada HMM.

Ya que se cuentan con 30 señas aisladas, o 10 señas compuestas, se cuentan con la misma cantidad de HMM. A cada uno de estos experimentos se le reporta la media y su respectiva desviación estándar.

Para el segundo paso que es identificar señas compuestas, se realiza un agrupamiento entre dos fonemas y tres fonemas articulados y realiza un nuevo entrenamiento de la misma manera que el anterior. El mismo proceso se realiza para las frases o señas compuesta.

4.4. Sistema de traducción

Para el sistema de traducción se crearon corpus en paralelo debido a que el sistema a implementar es un SMT el cual necesita dos corpus como se explicó en la sección 2.3. Para generar el corpus en paralelo se eligió un contexto académico, específicamente el punto de información de un centro de educación superior. Para la selección de las frases en español más comunes en un punto de información académico. Se realizó mediante la consulta de las preguntas más frecuentes en los centros de educación superior en roles de estudiante, aspirante, administrativo, egresado o externo; una vez definidas las frases, fue necesario ampliar el corpus. Esto se hizo añadiendo frases en español que son sinónimos de las frases ya existentes o que tienen la misma traducción en LSC. Todo esto con el objetivo de crear un sistema más robusto. Por lo tanto, se incluyen sinónimos de nombres, adjetivos o verbos y se modifica la sintaxis de las oraciones. Para contar con los dos corpus se realiza la traducción de las frases del corpus en español a glosas, está labor la realizó una persona experta en LSC.

Los corpus diseñados contiene un total de 517 frases. Algunos ejemplos del corpus generado se pueden apreciar en la tabla 1.

Español	Glosas
Necesito resolver una duda	Necesitar duda resolver
¿Hay dónde imprimir dentro de la universidad?	Imprimir ¿Dónde?
¿Dónde puedo ver el cronograma de las fechas de inscripción?	Cronograma inscribir fecha ver ¿Dónde?
¿Cuál programa requiere examen de admisión?	Examen admisión curso ¿Cuál?

Tabla 1: Ejemplo de frases del corpus en paralelo

A los dos corpus, tanto el de español como el de glosas, se les obtiene un modelo de lenguaje en bigramas y trigramas debido a que si se plantea un modelo con menos o más N-gramas no se tendría una cantidad de datos adecuada para realizar el entrenamiento del sistema de traducción. El esquema general del modelo de traducción se observa en la figura 25.

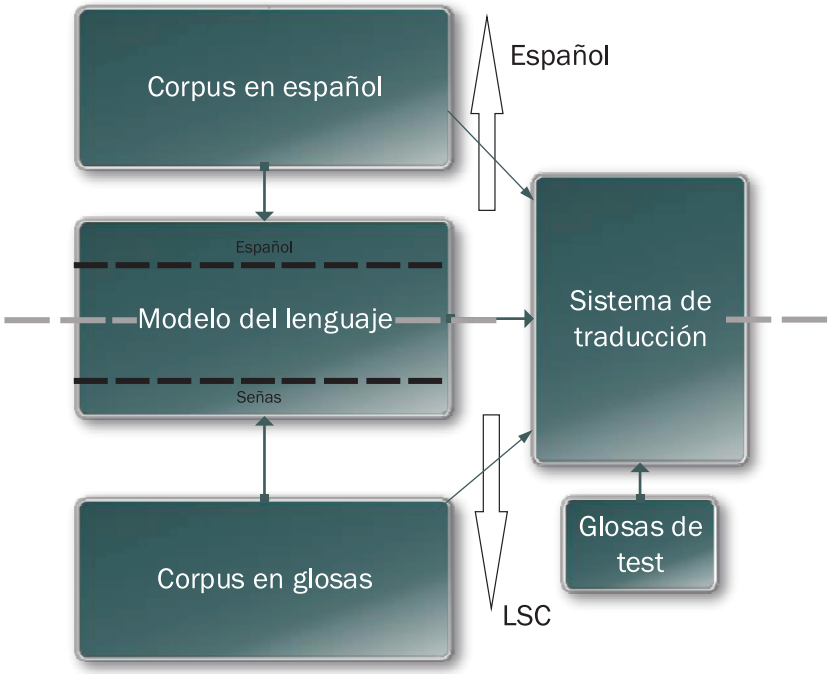


Figura 25: Modelo del sistema de traducción de glosas al español

Para obtener el modelo del lenguaje en glosas se utilizan bigramas y trigramas, como los de las tablas 2 y 3.

Idioma	Frase	Bigramas
Español	¿Cuál programa requiere examen de admisión?	¿Cuál programa - programa requiere - requiere examen - examen de - de admisión?
LSC	Examen admisión curso ¿Cuál?	Examen admisión - admisión curso- curso ¿Cuál?

Tabla 2: Ejemplo de bigramas del corpus creado

Idioma	Frase	Trigramas
Español	¿Cuál programa requiere examen de admisión?	¿Cuál programa requiere - programa requiere examen - requiere examen de - examen de admisión?
LSC	Examen admisión curso ¿Cuál?	Examen admisión curso - admisión curso ¿Cuál?

Tabla 3: Ejemplo de trigramas del corpus creado

En la tabla 4 se puede apreciar las características completas de los corpus utilizados para desarrollar este trabajo.

Descripción	*Corpus* en Español	*Corpus* en Glosas
Número total de frases	517	517
Número de preguntas simples	165	165
Número de preguntas compuestas	323	323
Número de afirmaciones	18	18
Número de saludos y despedidas	11	11
Longitud promedio de la frase	6.26	4.11
Número de palabras que contienen la frase más larga	13	9
Número de palabras que contienen la frase más corta	1	1
Número total de *tokens*	3237	2128
Número total de palabras	650	390
Palabra más repetida	de	dónde
Número de veces que aparece la palabras más repetida	238	132
Número de bigramas	1660	1277
Número de trigramas	2001	1633

Tabla 4: Corpus utilizados para generar el modelo de traducción de este trabajo

Una vez definidos los corpus se procede a entrenar el modelo de traducción. El modelo de traducción implementado se basa en un sistema estadístico de traducción que consta de los siguientes parámetros.

- Distorsión (D): permite reordenar la oración de entrada.
- Penalidad por frase (π): garantiza una buena traducción entre las palabras del lenguaje original y su traducción
- Penalidad por palabra (W): se encarga de que las traducciones no sean demasiado largas o cortas.
- Peso del modelo del lenguaje (LM): garantiza que la traducción sea fluida en su lenguaje original.

Para definir la traducción se realiza el cálculo de la probabilidad con la siguiente fórmula:

$$p(es|gl) = \pi(gl|es)^{peso\,de\,\pi} * LM(es)^{peso\,de\,LM} * D(es, gl)^{peso\,de\,D} * W(es)^{peso\,de\,W}$$

Donde

- $p(es|gl)$ es la probabilidad de una traducción al español dada una glosa.
- $\pi(gl|es)$ probabilidad de la tabla de traducción de las frases de español a glosas.
- $LM(es)$ modelo del lenguaje de español.
- $D(es, gl)$ distorsión entre el español y las glosas.
- $W(es)$ penalidad por frases.

Para realizar el sistema de traducción se parten las 512 frases en 10 particiones, donde se entrena 10 modelos cada uno con el 90 % de las frases y se valida con el 10 %, esta selección se realiza de manera aleatoria, a cada uno de estos modelos se les calcula el BLEU y el WER con el fin de determinar la calidad del sistema de traducción, de igual manera se realizó un cambio en los diferentes pesos del

modelo de traducción con el fin de ajustarlo y así conseguir una mejor traducción. Para los valores de peso de π, peso de LM y peso de D se varían desde 0.1 a 1. Y para el peso de W se varían entre -3 a 3. Para cada uno de estos experimentos se reporta la media y su desviación estándar, tanto para BLEU y WER.

5. ANÁLISIS Y RESULTADOS

En este sección se muestran todos los experimentos realizados con sus respectivos resultados. Se divide en dos secciones, en la primera sección 5.1 se muestran los diferentes descriptores, métodos, agrupamientos, entrenamiento y validaciones que se hicieron para realizar el reconocimiento de las señas. En la sección 5.2 se muestran los sistemas de traducción y los resultados que se obtuvieron al variar los parámetros de cada modelo con el fin de obtener un modelo sintonizado, que genere una mejor traducción. En todos los experimentos realizados se reporta la tasa de acierto y su respectivo intervalo de confianza.

5.1. Reconocimiento de señas

Para el reconocimiento de las señas se plantearon diferentes descriptores que permitieran reducir la dependencia de los datos respecto a la ubicación de la persona y de sus medidas antropométricas. La forma en que se extrajo la información se explicó en la sección 4.3. Utilizando las características extraídas se realizó un método de agrupamiento con el fin de determinar los fonemas articulados, para esto se utilizó el coeficiente de silhouette, todo esto se realiza con el fin de imitar los estudios presentes del lenguaje de señas y poder encontrar la matriz de articulación y la matriz segmental.

Uno de los agrupamiento realizado en el trabajo se presenta figura 26. En este fue necesario reducir a 3 dimensiones con el fin de poder graficarlo.

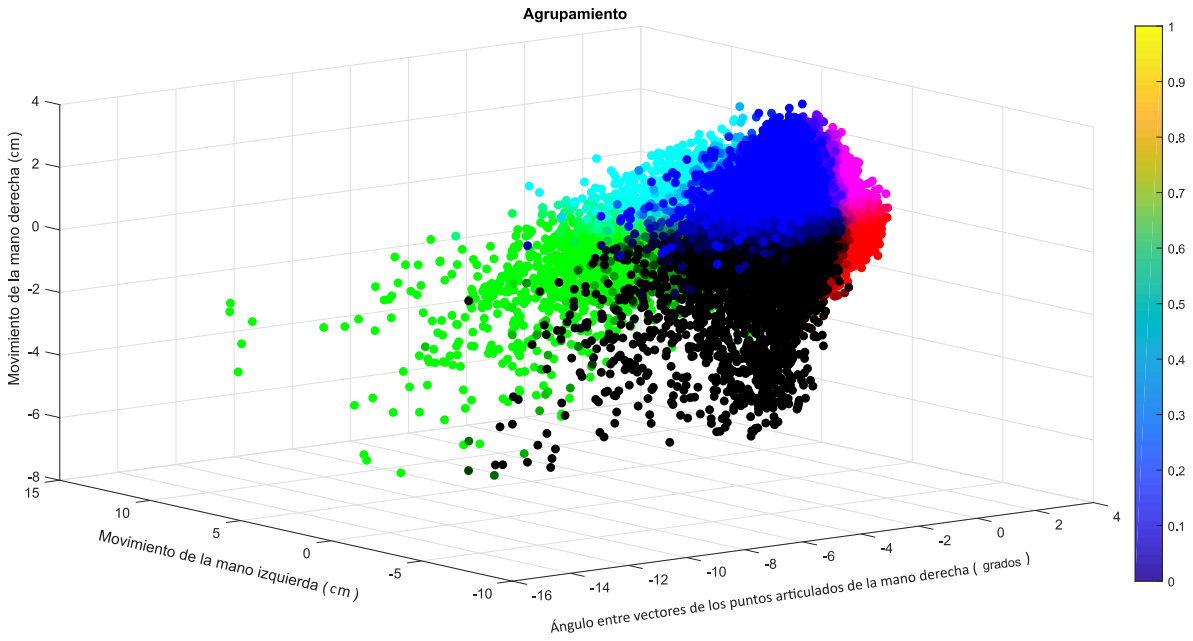


Figura 26: Agrupamiento de la matriz de características

Los datos que se dibujan en la figura 26 son tres de las características que se explicaron en la sección 2.2. Ahora se mostraran los resultados obtenidos para cada uno de los métodos propuestos en la sección 2.2

5.1.1. Señas aisladas

Para cumplir con el objetivo de definir el diccionario de fonemas articulados es necesario validar cada uno de los métodos explicados en la sección 4.3. A continuación se mostraran los resultados para cada uno de los descriptores utilizados:

Cambio de punto de origen Con el primer descriptor lo que se buscó fue eliminar la dependencia respecto a la ubicación donde el actor realizara la seña como se explicó en la sección 4.3.1. El primer paso fue agrupar los datos extraídos utilizando K-means y calcular su respectivo coeficiente de silhouette. Todo esto se realiza para determinar los fonemas articulados y poder determinar en cuantas partes se divide las señas con las que se cuenta.

En la figura 27 se presenta el coeficiente de silhouette con el primer descriptor propuesto, se puede observa que el mejor agrupamiento se encuentra en 3 *cluster*.

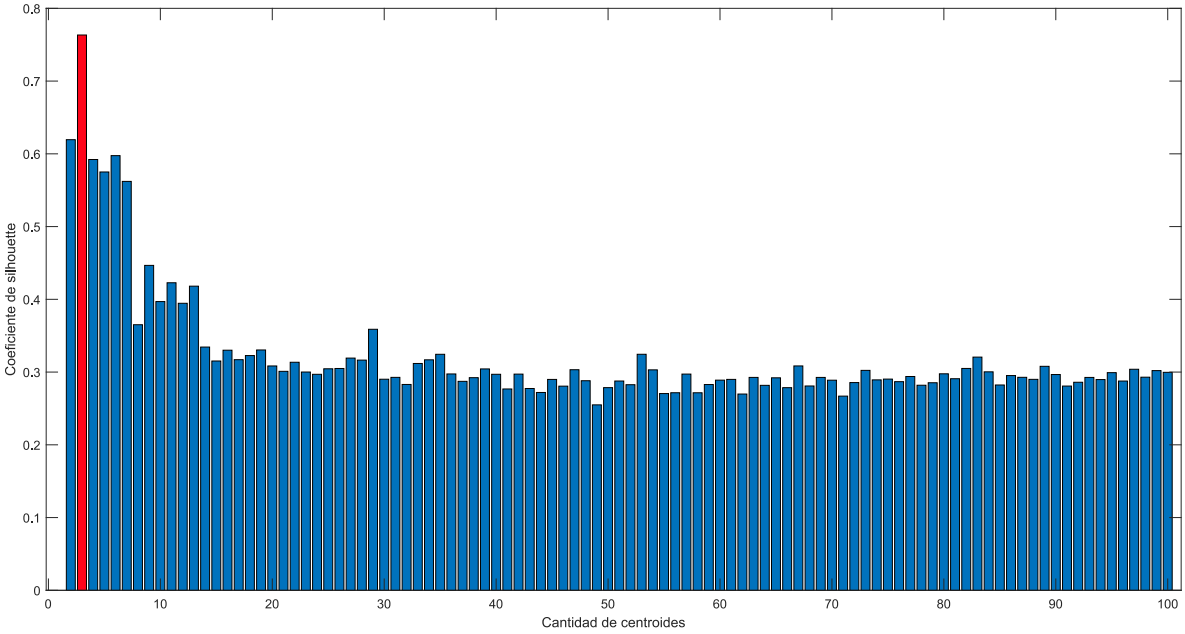


Figura 27: Coeficiente de silhouette con descriptor de punto de referencia

Dado que el coeficiente de silhouette indica que el mejor agrupamiento se encuentra en el rango de 3 a 5 grupo, dentro de las metodologías propuestas en [14, 13] no se muestra claramente como se realizó la etiqueta de la base de datos y no se muestra la cantidad de grupos que se deben seleccionar. Por lo tanto con esto se busca realizar un mejor análisis con el fin de definir los micromovimientos o fonemas articulados, es decir, los movimientos mínimos para crear las diferentes señas como se realiza con los lenguajes hablados.

Con el fin de corroborar los datos obtenidos por medio del coeficiente de silhouette, se realiza un entrenamiento de 30 HMM, una para cada seña aisladas y se valida realizando una validación en 6

partes, se entrena con 5 y se valida con uno, este proceso se realiza con cada uno de los descriptores descritos en la sección 4.3, en cada una de las figuras se indica con un color rojo el mejor agrupamiento según el coeficiente de silhouette, como se puede observar en la figura 28. De estos datos se aprecia que el mayor porcentaje de acierto es de $37.78 \pm 4.67\,\%$, este resultado se presenta cuando se cuenta con grupos variando del resultado entregado por silhouette, también se observa que después de 10 grupos tienden a tener un comportamiento similar respecto a la clasificación de señas.

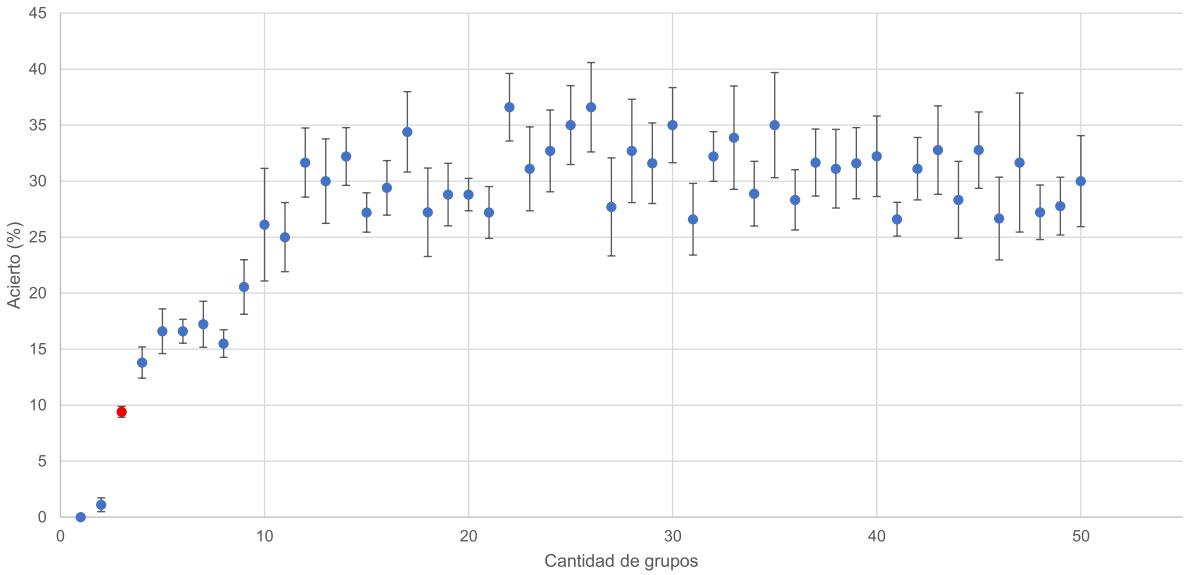


Figura 28: Resultados de descriptores de distancias con referencia a un punto

Cambio de punto de origen y normalización de la magnitud. Ahora se plantea el uso de un nuevo descriptor que reduzca la dependencia respecto a la ubicación donde el actor realizara la seña y sus medidas antropométricas como se explicó en la sección 4.3.1. Al igual que el punto anterior, los datos extraídos se agrupan por medio de K-means y se realiza el cálculo del coeficiente de silhouette. Por medio de esta medida se intentaron determinar los fonemas articulados y conocer en cuantas partes se divide las señas con las que se cuenta para el descriptor propuesto. En la figura 29 se aprecia que nuevamente el mejor agrupamiento según el coeficiente de silhouette se presenta en 3 *cluster*.

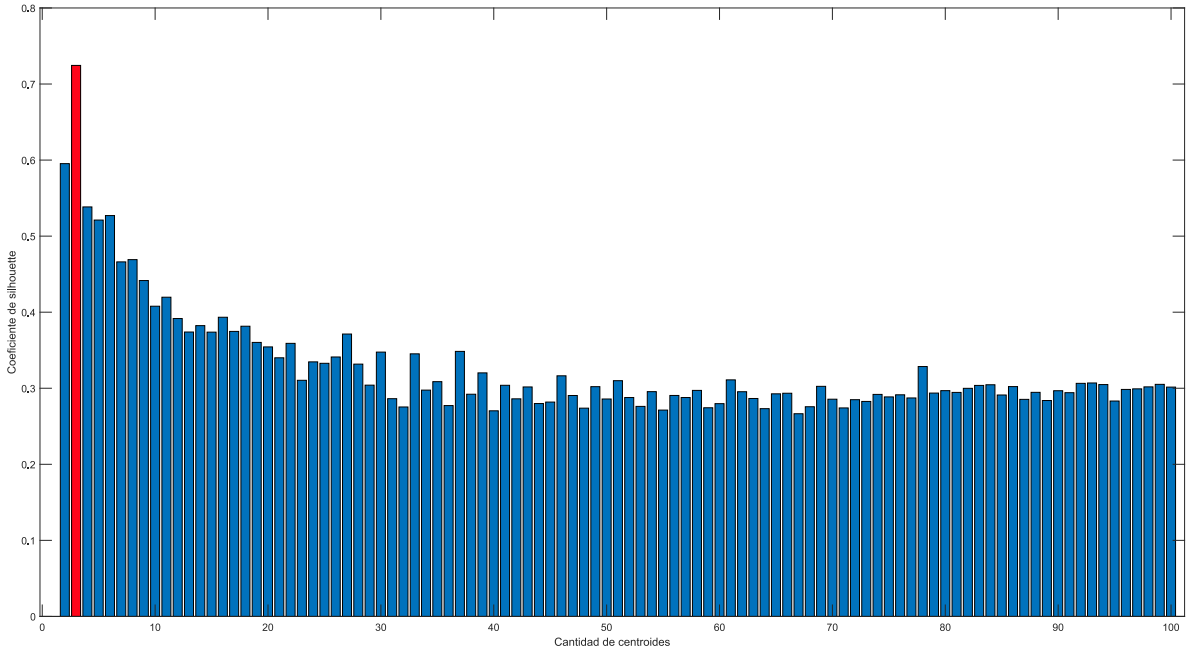


Figura 29: Coeficientes de silhouette de cada una las características extraídas

Nuevamente para determinar si el valor entregado por el coeficiente de silhouette genera la mejor clasificación se realiza nuevamente el entrenamiento de HMM, los datos que se obtienen se observan en la figura 30, demostrando que cuando se normalizan respecto a una medida del cuerpo los resultados mejoran ya que se pasa de tener una tasa de acierto del $37.78 \pm 4.67\,\%$ a $36.6 \pm 3.99\,\%$, demostrando que la metodología propuesta mejora los resultados a nivel de confianza en el resultado, sin embargo, al igual que en el caso anterior, la mejor clasificación no se encuentra en el punto que entrega el coeficiente de silhouette, sino en el agrupamiento que presenta 22 *cluster*.

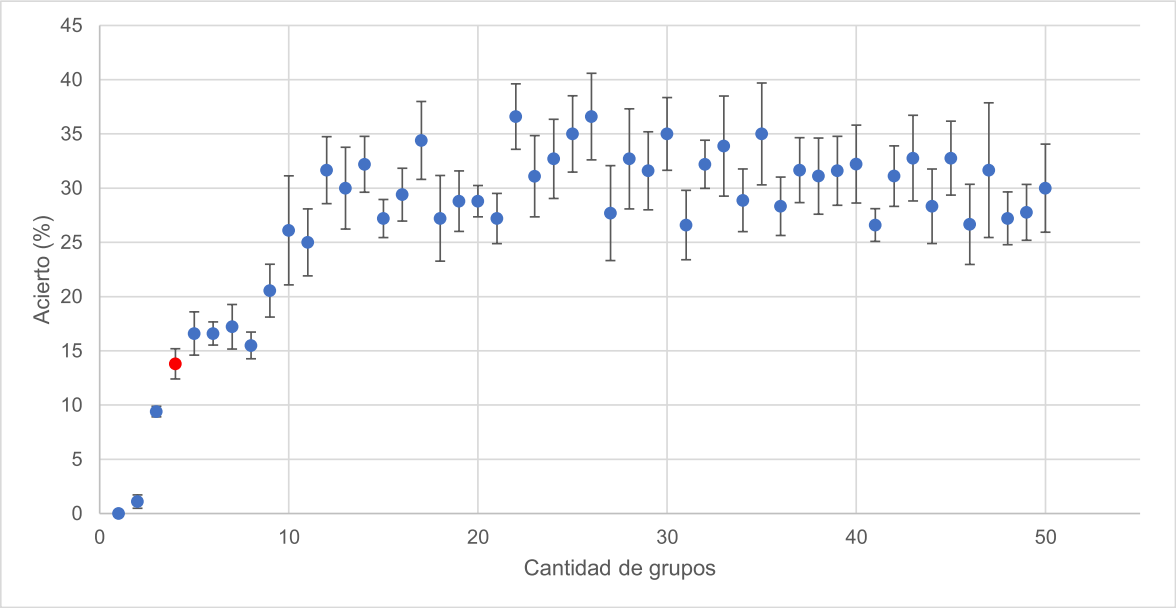


Figura 30: Resultados obtenidos con descriptor de distancia y referenciado

Ángulos normalizados y distancias. En este punto se realiza un análisis de correlación con el fin de eliminar datos redundantes y así poder reducir el costo computacional de los algoritmos, ya que se cuentan con 24 características que representan los 8 ángulos de los vectores que se explicaron en la sección 4.3.1. El análisis de correlación se muestra en la figura 31, en los cuadros blancos se muestran los descriptores relacionados con los hombros, en el cuadro rojo se encuentran los descriptores de la mano derecha y en el negro los de la mano izquierda, adicional en el rectángulo naranja se encuentran las distancias entre los dedos se muestran los ángulos relativos de los diferentes punto articulados respecto a la nueva coordenada de origen.

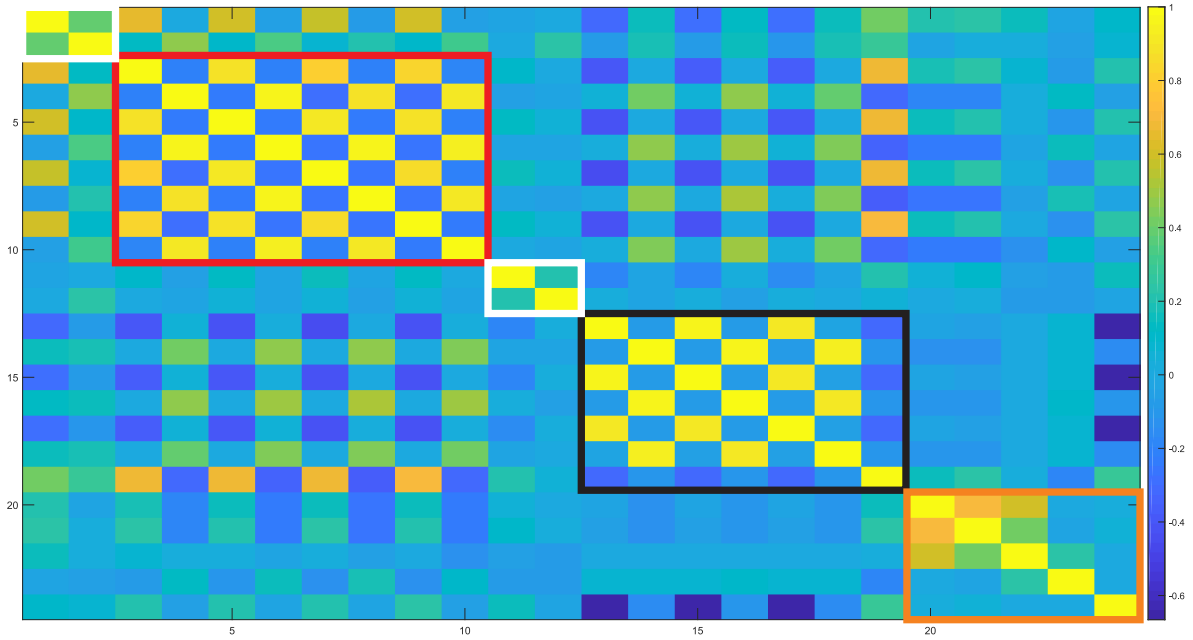


Figura 31: Correlación entre descriptores utilizados

Una vez se analizada la gráfica de correlación de los descriptores obtenidos se puede apreciar que la mano derecha y sus respectivos puntos al igual que la manos izquierda se encuentran fuertemente relacionados, por lo tanto se realizan dos experimento. El primero con todos los datos como se puede observar en la figura 32 obteniendo un resultado máximo de eficiencia del 36.6 % y el segundo con 11 características, se eliminan los datos redundantes, quedando una característica por cada mano, esta característica representa el movimiento que realiza cada una de las manos, los datos de este segundo experimento se muestran en la figura 33.

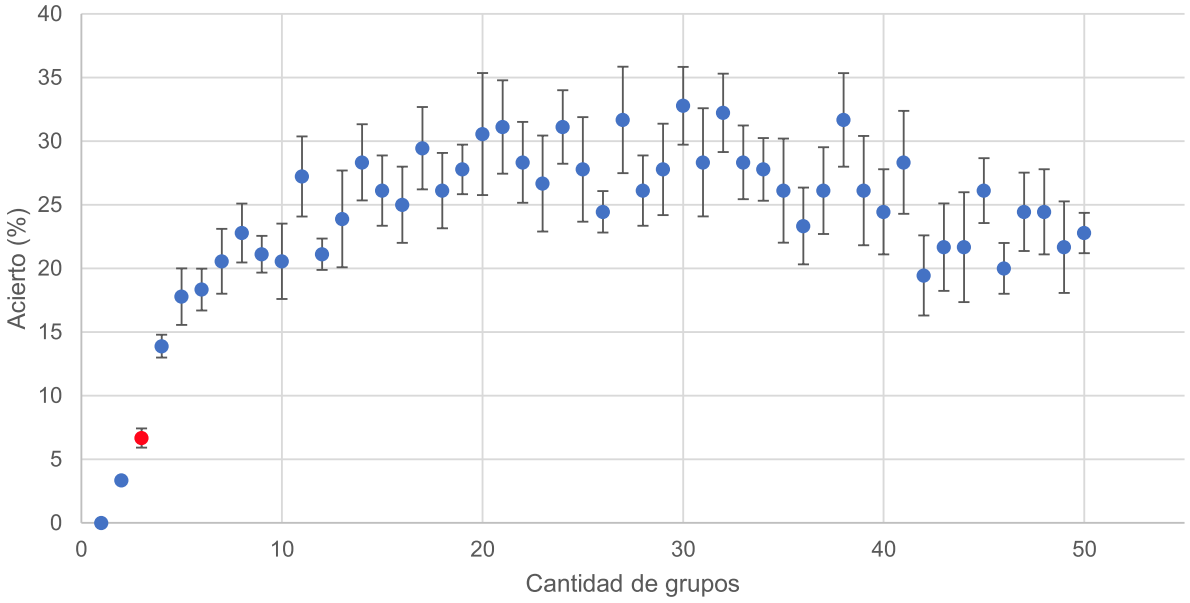


Figura 32: Resultados obtenidos al utilizar el descriptor propuesto

Los resultados al reducir la cantidad de características muestran un leve aumento en la tasa de acierto, ya que paso de 34.6 % a 36.5 %, con lo que se demuestra que no es necesario contar con las 24 características calculadas. De la figuras 32 y 33, se puede apreciar que al igual que los descriptores anteriores el coeficiente de silhouette difiere de la mejor clasificación, la cual se encuentran para el primer caso en 30 *cluster* y para el segundo caso en 36 *cluster*.

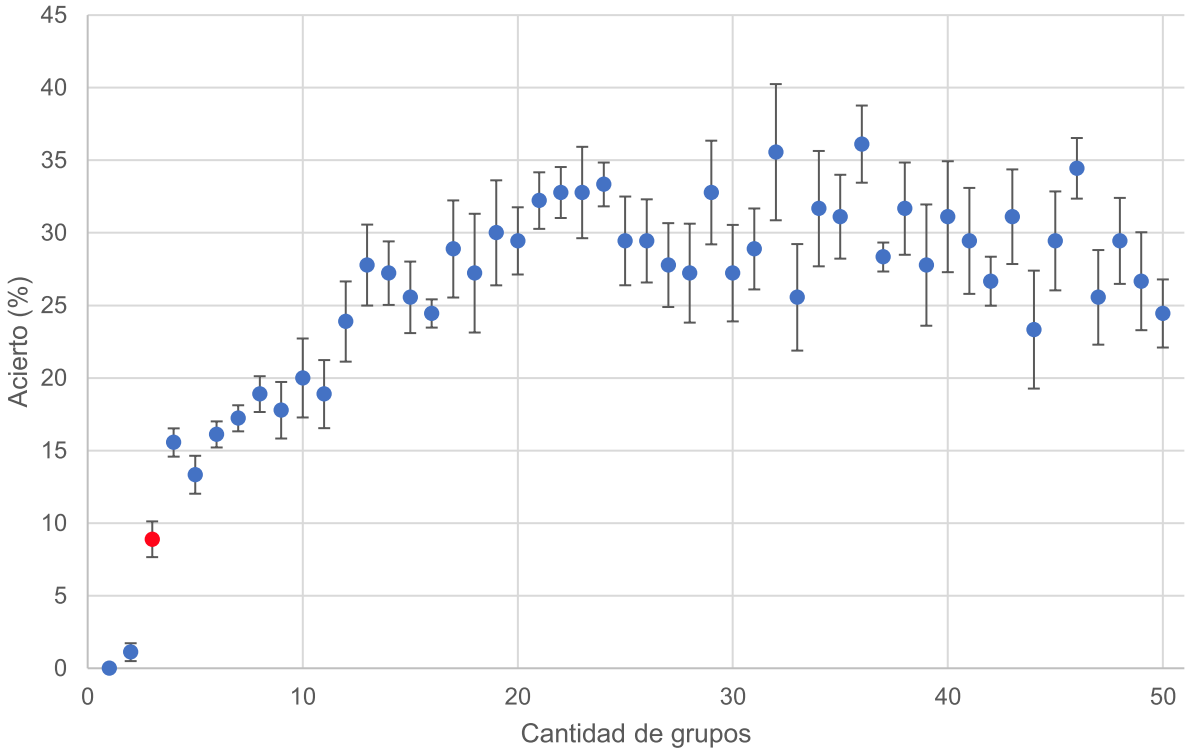


Figura 33: Resultados obtenidos al utilizar el descriptor propuesto y reducción de características

La metodología planteada presenta resultados mayores al 30 %, los resultados obtenidos no se pueden mejorar debido a las pocas repeticiones que se tienen de una misma seña. Aunque no es posible compararse con otros trabajos del estado del arte, se han seguido metodologías empleadas por otros, con variaciones propias llegando a los resultados mostrados anteriormente. Por otro lado, se destaca que con este trabajo se tuvieron en cuenta signos de la lenguaje de señas colombiana, es decir, que presentan y no señas estáticas como lo realizaron en la Universidad Libre de Colombia [15].

También las metodologías que presentan mejores tasas de eficiencia trabajan con grandes flujos de datos, ya que la extracción de características la realizan a nivel de imágenes tanto de color como de profundidad e ignoran los puntos articulados. Demostrando que este trabajo puede ser un nuevo punto de partida para trabajar. Por otro lado se debe tener en cuenta que las metodologías planteadas en el estado del arte normalmente no se ajustan a gestos del lengua de señas colombiano, y que el trabajo que más se acerca a esté se realizó en la Universidad de los Andes [16] y presenta una base de datos pequeña y que sus mejores porcentajes de acierto son al nivel de señas estáticas.

El otro punto clave de este trabajo fue la definición fonemas articulados, ya que en muchos trabajos realizados esto no se definen, por lo tanto no se realiza una compresión desde el modelo de la lengua sino que se enfocan en un análisis a nivel de clasificación. Por lo tanto intentar definir los fonemas articulados genera un avance a futuro sobre los trabajos de este tipo y poder definir de una manera sistémica los movimientos básicos que componen una seña. Aunque el punto de partida fue utilizar el coeficiente de silhouette, este no presenta una verdad absoluta, ya que en este trabajo la mejor cohesión entre datos y la separación de los grupos se presento entre 3 a 5 grupos sin embargo la mejor

clasificación se tuvo en el rango entre 20 a 36 grupos, lo cual demuestra que se puede comportar como un lenguaje hablado como el español que presenta 30 fonemas o sonidos mínimos que conforman cada una de las palabra que compone este tipo de lenguaje.

5.1.2. Señas compuestas

Para el reconocimiento de señas compuestas se utiliza la misma metodología que se planteó para señas aisladas, en este caso la base de datos esta compuesta por 10 frases completas y se repite nuevamente su clasificación con los descriptores del estado del arte y con el descriptor propuesto.

Para el primer descriptor, los datos se pueden observar en la figura 34, en este caso se observa un acierto de 66.66 % con diez grupos.

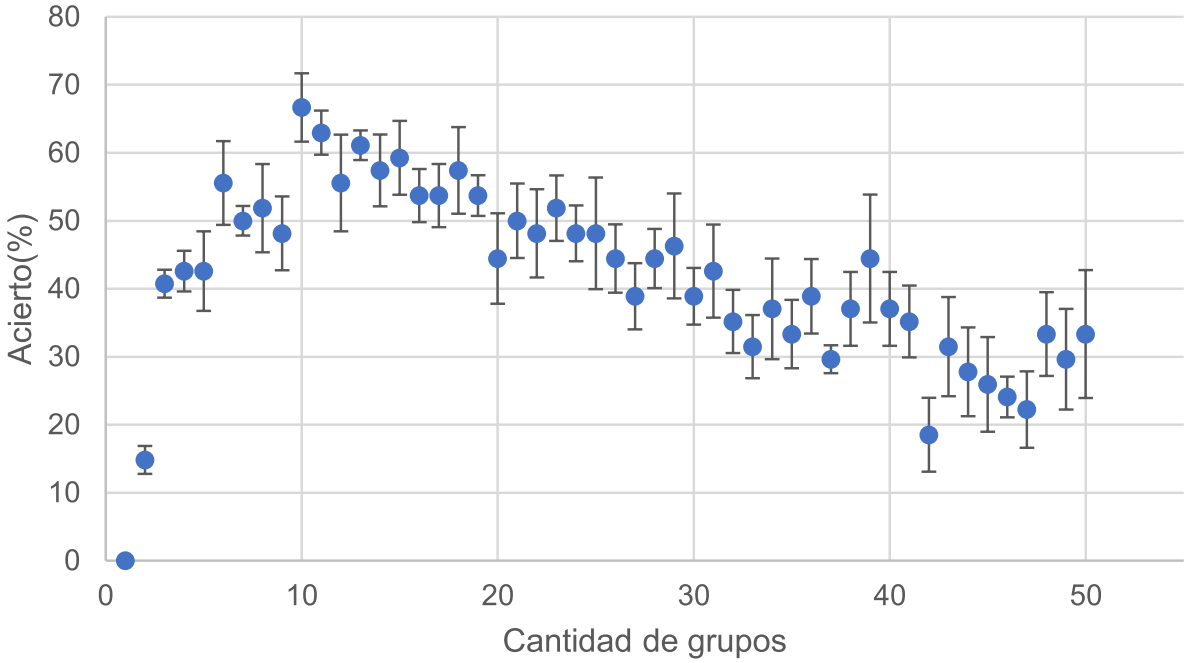


Figura 34: Resultado de frases compuestas con descriptor uno

Nuevamente se procede a realizar el proceso de normalización dividiendo los datos por la distancia entre el punto medio del cuerpo y la cabeza, los resultados se observan en la figura 35, en esta se observa que el mejor resultado se encuentra al utilizar nueve grupos con un porcentaje de acierto de 64.81 % reduciendo en 2 % el resultado obtenido con el anterior descriptor.

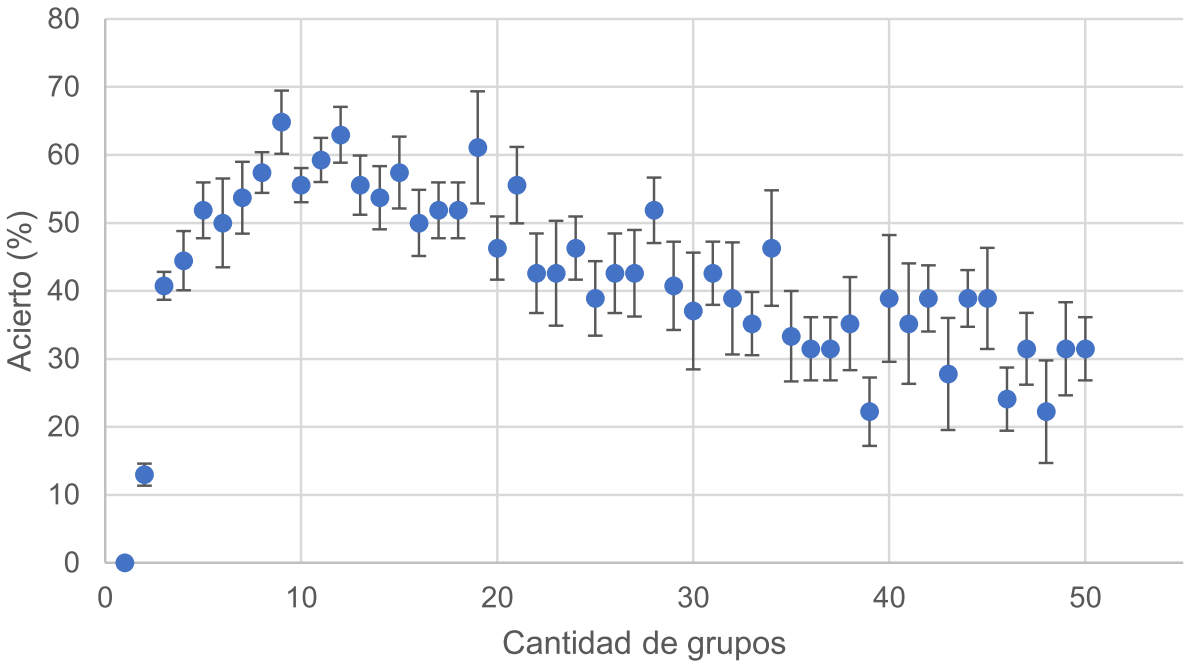


Figura 35: Resultado de frases compuestas con el segundo descriptor

Por último, se utiliza la metodología planteada que permite disminuir las características de los datos como se mostró en 5.1.1. Los resultados se pueden observar en la figura 36, obteniendo una tasa de acierto del 72.22 % para el agrupamiento de 10 y 11 *cluster*, demostrando que la metodología propuesta mejora los resultados frente a los descriptores que se plantean en el estado del arte para la identificación de señas compuestas.

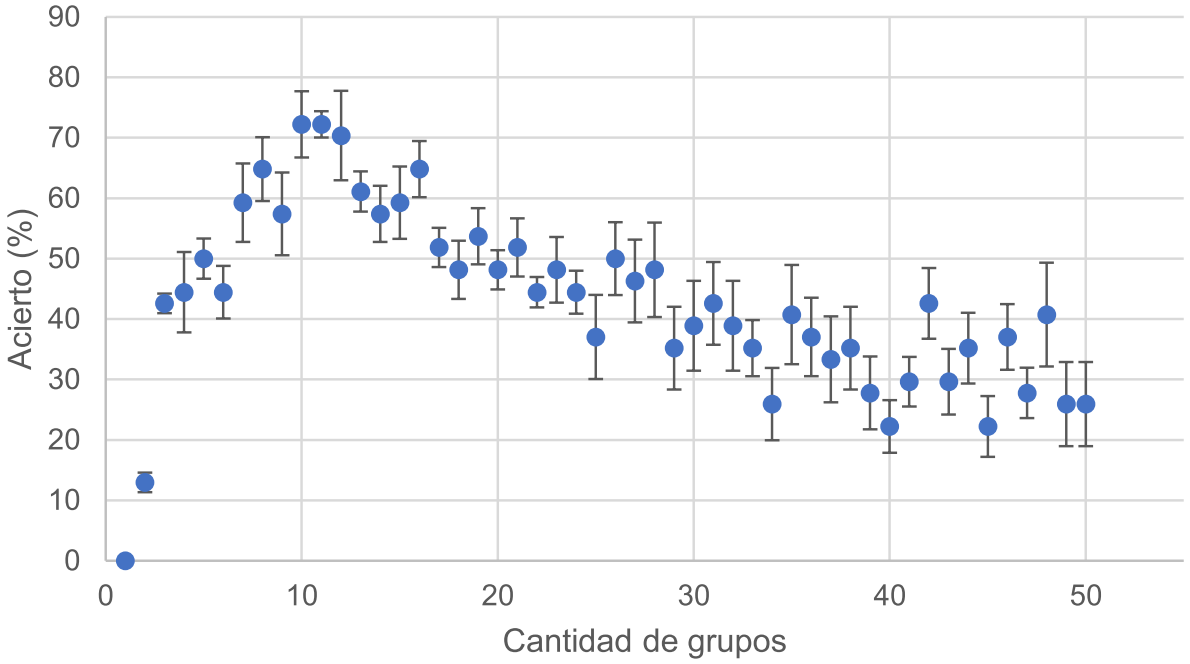


Figura 36: Resultados con metodología planteada y reducción de características

De los resultados mostrados en esta sección, se puede observar que aumentaron significativamente frente a los que se presentan en las señas aisladas, demostrando que la metodología propuesta sirve para señas aisladas y compuestas a diferencia de las que se plantean en [16, 15, 53, 14, 54] ya que en en estos trabajos se identifican señas aisladas y en algunos de ellos solo estáticas. Cabe resaltar que el método que se planteó para realizar el reconocimiento de señas compuestas, es parecida a la utilizada para el reconocimiento de frases en los lenguajes hablados, demostrando que la metodología propuesta de ser tratado como un lenguaje natural puede ser válida. Se espera que por medio de este proyecto se pueda contribuir a un mejor entendimiento de este tipo de lenguaje con el fin de ayudar a futuro a la inclusión de las personas que presentan sordera parcial o total y a la comunidad como personas activas de la mismas.

5.2. Validación del modelo de traducción

Una vez determinada la seña realizada se procede a una segunda parte la cual corresponde en traducir de glosas a español, por tanto se implementa un sistema de traducción estadístico como el explicado en la sección 4.4. Para realizar esta etapa se dividen los datos de manera aleatoria con el fin de realizar una validación cruzada, se toma el 90 % del corpus para crear un modelo de traducción y se valida con el 10 %, este proceso se realiza 10 veces utilizando modelos de lenguaje basados en bigramas y trigramas, la división realizada se observa en la tabla 5

Modelo	Número de 1-grama	Número de 2-grama	Número de 3-grama
Modelo 1	634	1536	1832
Modelo 2	634	1536	1832
Modelo 3	632	1544	1832
Modelo 4	625	1525	1821
Modelo 5	632	1540	1841
Modelo 6	631	1543	1840
Modelo 7	632	1540	1841
Modelo 8	629	1515	1802
Modelo 9	631	1543	1840
Modelo 10	632	1540	1841
Promedio	632	1536	1832

Tabla 5: Cantidad de unigramas, bigramas, trigramas resultantes de la generación de cada uno de los modelos

Para cada uno de los modelos se realiza una búsqueda de manera heurística para sintonizar los pesos del modelo de traducción que permita obtener la mejor traducción. Se toman como punto de partida los definidos por el *toolbox* utilizado para implementar el sistema de traducción [55].

5.2.1. Modelos utilizando bigrama

Para cada uno de los modelos que se obtienen por medio de bigramas se les realiza un cambio en los parámetros del peso del modelo del lenguaje y la distorsión con el fin de encontrar los pesos que generan una mejor traducción, solo se trabaja con estos dos parámetros ya que los demás para este caso no presentan un cambio significativo. Es de aclarar que los valores que se van a modificar se encuentran normalizados. Por medio de BLEU y WER se mide la calidad en la traducción los resultados se pueden apreciar en la figuras 37 y 38. Cada figura realizada por el WER reporta el *Word Accuracy* W_{ACC}, esto con el fin de poder comparar las dos métricas utilizadas respecto a la cantidad de aciertos. Para las dos figuras se toma un valor de distorsión de 0.3, penalidad por frase de 0.2 y penalidad por palabra -1, valores recomendados en [55].

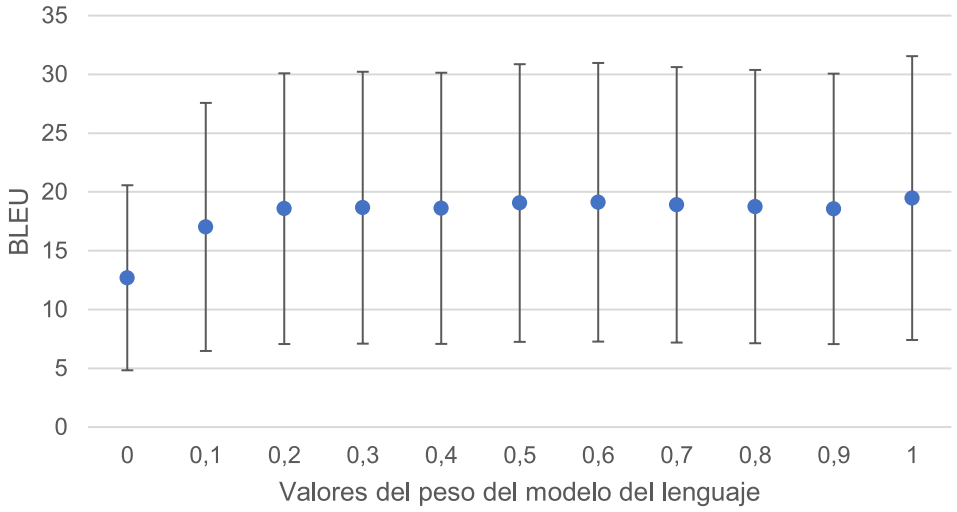


Figura 37: Validación del modelo por medio de BLEU variando el peso del modelo del lenguaje

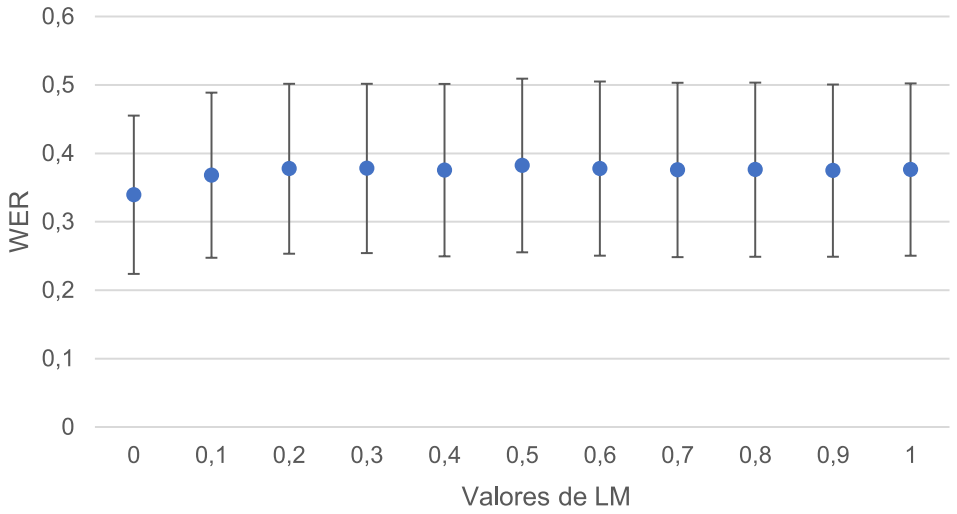


Figura 38: W_{ACC} del modelo por medio variando el peso del modelo del lenguaje

De los anteriores resultados se observa que la traducción es mejor cuando el peso del modelo del lenguaje es cercano a uno, por lo tanto se toma un valor de 1 para este parámetro y se realizó un cambio en el parámetro de distorsión los resultados se pueden observar en las figuras 39 40. En estas se observan que para BLEU si se ve afectado mientras que para WER permanece casi constante. Lo mismo sucede al cambiar la distorsión del modelo del lenguaje, como se puede apreciar en las figuras 39 y 40

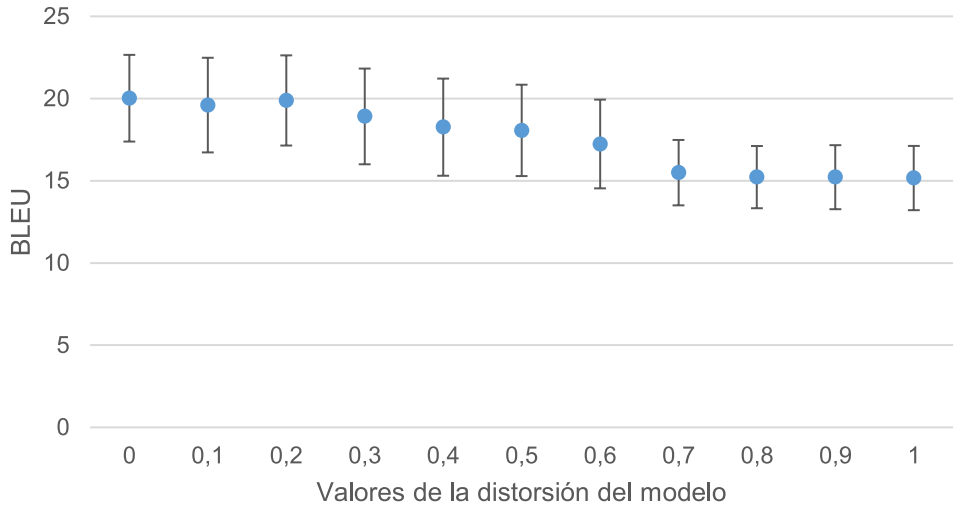


Figura 39: Validación del modelo por medio de BLEU variando la distorsión

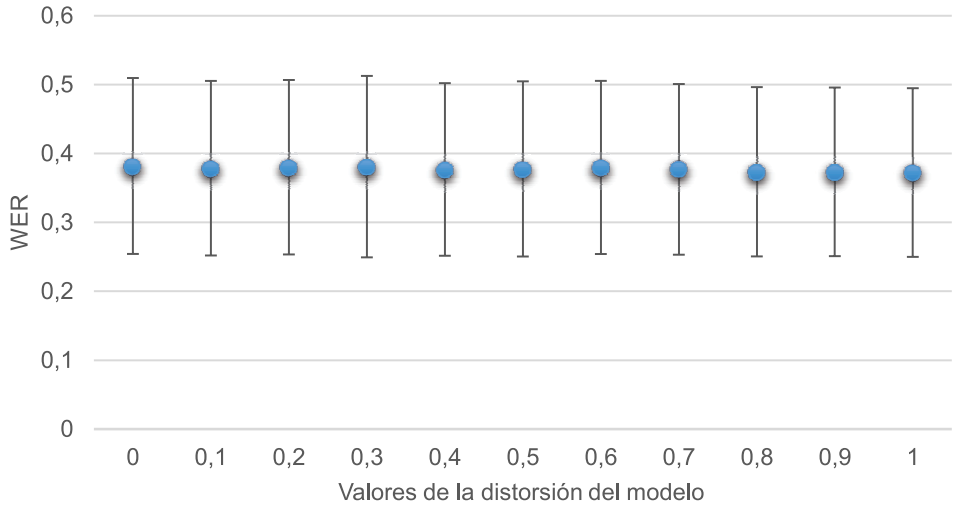


Figura 40: W_{ACC} del modelo por medio variando el peso del modelo del lenguaje

De los anteriores gráficas se observa que WER presenta un gran gran valor, lo que significa que la traducción no es buena. Sin embargo a la hora de realizar un análisis más exhaustivo se aprecia que la métrica de WER no es confiable, ya que no considera la interpretación sino que compara textualmente el texto de salida y la traducción, por lo tanto cualquier cambio pequeño puede presentar grandes errores, en la tabla 6 se presenta un ejemplo de este tipo de problema. En esta tabla se puede observar que una traducción realizada por el sistema el valor de WER indica un error del 57.14 % entre la traducción realizada y su referencia en español. Pero por otro lado, BLEU entrega un valor de acierto del 75 % indicando que la traducción presenta un buen comportamiento.

Glosa	contraseña correo universidad recuperar ¿cómo?
Traducción	¿cómo puedo recuperar mi contraseña del correo institucional
Español	¿cómo recupero la contraseña del correo institucional?
WER	57.14 %
BLEU	75 %

Tabla 6: Comparación entre BLEU y WER como métricas de traducción

5.2.2. Modelos utilizando trigrama

El proceso realizado para bigramas se repite para trigramas y de nuevo se empieza variando el peso del modelo del lenguaje y se asigna un valor de distorsión de 0.3, penalidad por frase de 0.2 y penalidad por palabra -1.

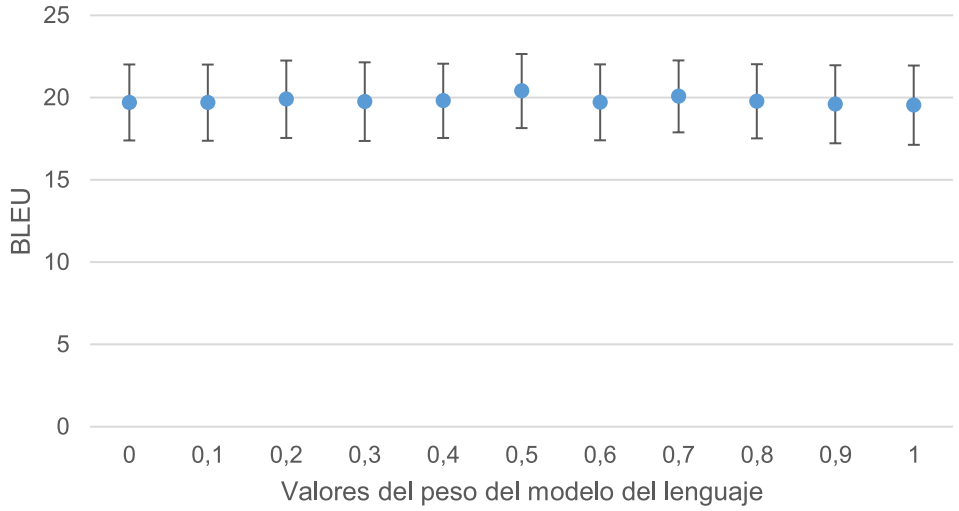


Figura 41: Validación del modelo por medio de BLEU variando el peso del modelo del lenguaje

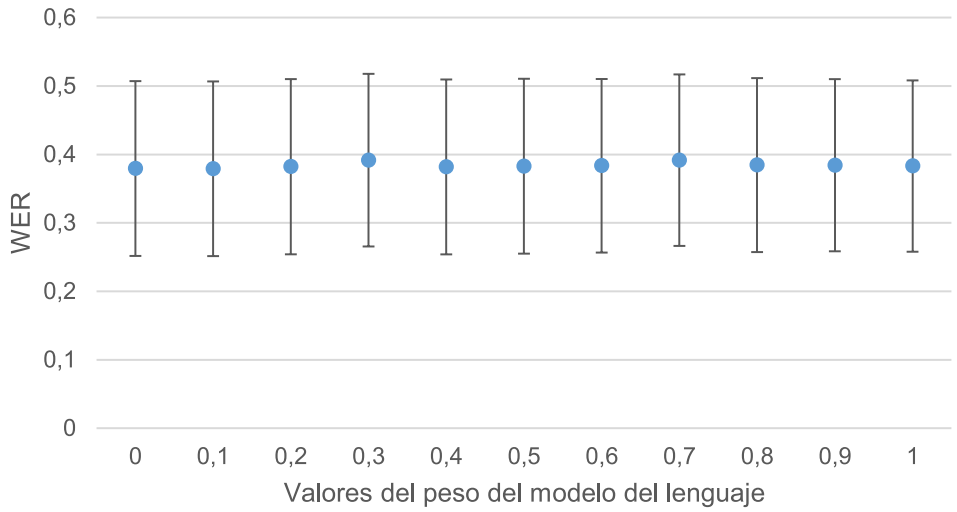


Figura 42: W_{ACC} del modelo por medio variando el peso del modelo del lenguaje

De los anteriores resultados, se observa que el cambiar el peso del modelo del lenguaje no genera ningún cambio para los sistemas de traducción creados. Los resultados se pueden observar en las figuras 41 42.

De los sistemas de traducción implementados se puede apreciar que los mejores modelos presentan un BLEU de 21.56 ± 4.3 y un W_{Acc} de 39.88 ± 23.64.

Se puede apreciar que los sistemas de traducción que se implementaron presenta un WER de 60.2 %. Esto se debe a que WER busca una comparación exacta entre palabras, dejando a un lado el significado de las mismas. Esto se puede eliminar utilizando una mejor métrica de medida como lo es BLEU, la cual permite una mejor medida de las traducciones realizadas.

Se esperaba un mejor resultado en los sistemas de traducción ya que se utilizan lenguajes que de manera escrita son similares, sin embargo esto no sucede, ya que no se tiene un corpus amplio de glosas y palabras que permita realizar una mejor traducción. Cabe resaltar que si el sistema se valida con una de las glosas que fue entrenado se obtiene un BLEU cercano al 75 %. Comparado el sistema desarrollado con [48] se puede apreciar que los resultados son similares, por lo tanto se puede comprobar que el sistema desarrollado es comparable con otros trabajos del estado del arte.

Dado que este es uno de los primeros trabajos que se implementan en el área, se hace necesario determinar si las métricas utilizadas para la evaluación de los sistemas de traducción de lenguas escritas presentan la misma validez que las métricas empleadas e los sistemas de traducción de una lengua ágrafa.

5.3. Validación del sistema

Con el fin de validar todo el sistema es necesario fusionar el sistema de traducción con el sistema de reconocimiento como se puede observar en la figura 43

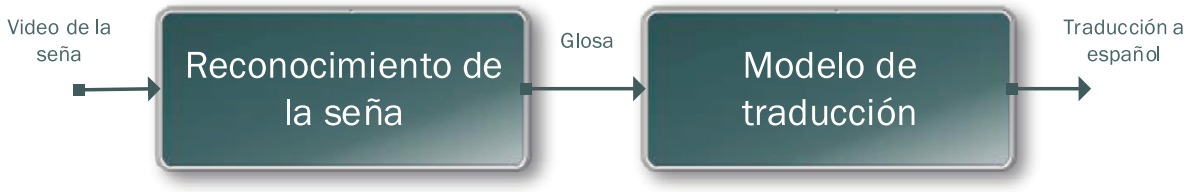


Figura 43: Sistema de reconocimiento y traducción

Para realizar el análisis completo se parte de los mejores resultados de señas aisladas, los cuales fueron obtenidos con el descriptor propuesto obteniendo un porcentaje de acierto de 72.22 %. Para validar el sistema se ingresan las diez frases de los diferentes actores y se observa si la traducción corresponde realmente a lo deseado. Cabe resaltar que tanto el sistema de traducción como el sistema de reconocimiento no supera el 20 % de acierto. Los resultados obtenidos se observan en la figura 44, obteniendo un porcentaje de acierto 17.33 % en 10 y 11 grupos.

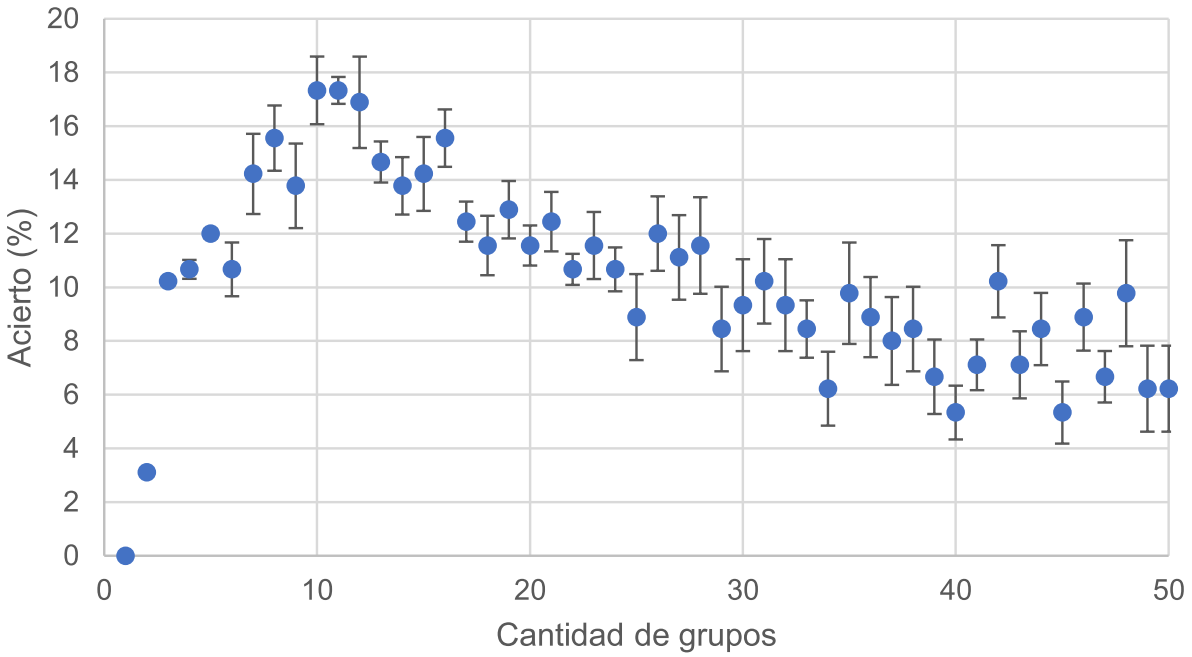


Figura 44: Resultados al validar todo el sistema

De los resultados obtenidos al validar todo el modelo, se puede apreciar que su eficiencia es baja, sin embargo es de aclarar que este tipo de análisis no se realiza normalmente no son trabajados en [13, 48, 16, 15, 53] y sencillamente se quedan en el reconocimiento de la seña, o en la evaluación del sistema de traducción como es el caso de [41] que se encargan de evaluar todo el sistema de traducción pero no del reconocimiento, por lo tanto se puede observar que el sistema propuesto aunque presenta bajo porcentaje de acierto se acerca a un sistema real que permita la traducción simultanea entre lenguas escritas y lengua ágrafa.

6. CONCLUSIONES

En este capítulo se muestran las conclusiones del trabajo realizado. De manera general se puede concluir que:

- Se desarrolló y se evaluó una metodología para la caracterización de imágenes en el reconocimiento de la lengua de señas colombiana y su traducción al español. Esta metodología permitió además observar cuáles eran los mejores descriptores mecánicos utilizados en la caracterización de los movimientos corporales que se hacen en lengua de señas.

- Se creó una base de datos anotada compuesta por imágenes y videos de personas sordas e intérpretes de la lengua de señas colombiana. Esta base de datos, la primera en su tipo en Colombia, se planeó y construyó a partir de interacciones propias de un sistema de información en una institución de educación superior. Esta se caracteriza por su heterogeneidad en la selección de los actores, y por servir como punto de partida para futuros estudios en el campo de procesamiento de lenguaje natural.

A continuación, y de manera más especifica, en 6.1 se presentan las conclusiones respecto al sistema de reconocimiento automático de señas; y en 6.2 se presentan las conclusiones respecto al sistema de traducción.

6.1. Reconocimiento de señas

- Se implementó y validó una metodología para el reconocimiento automático de la lengua de señas colombiana utilizando técnicas de aprendizaje de máquina que hagan uso del diccionario de fonemas articulados creado para este trabajo.

- Se determinó un espacio vectorial de representación para las características que se obtienen a partir de los puntos articulados entregados por un sensor de profundidad. El análisis de estas características determinó la necesidad de eliminar la dependencia inter-subjetiva de las señas respecto a las medidas corporales de cada persona. Por esta razón se utilizaron descriptores normalizados, tales como ángulos entre los vectores a analizar (ver figura 33), los cuales permiten no solo eliminar la dependencia ya mencionada, sino también determinar con detalle la posición de los dedos, ya que en la gran mayoría de las señas la información se encuentra contenida en ellos.

- Se formuló una metodología para la creación de un diccionario de «fonemas» articulados utilizando técnicas de agrupamiento supervisado. El número de fonemas se determinó a partir de distintas métricas que dieran cuenta de la cohesión entre datos de un mismo *cluster* y la separación de los mismos. Los experimentos mostraron que los mejores resultados en clasificación se obtuvieron empleando entre 20 y 33 fonemas articulados (ver figura 36). Estos fonemas permiten representar la información de las 199 señas, que contiene la base de datos, en un conjunto mínimo de movimientos. Cabe destacar que en distintas aplicaciones del procesamiento del lenguaje natural, como son los sistemas de reconocimiento automático del habla y sistemas de traducción automática suelen emplearse un número similar de fonemas en la voz.

- El análisis de correlación entre los distintos descriptores determinó que existe una alta densidad de información redundante entre algunos de ellos, específicamente entre los puntos articulados de las manos. Los experimentos demostraron que el empleo de un solo punto articulado en las manos (el punto correspondiente a la palma de la mano) es suficiente en la caracterización y reconocimiento de señas (ver figura 31). Este análisis de correlación y la posterior reducción del espacio de características que de él se desprende, permiten reducir el costo computacional de la implementación de la metodología aquí presentada.

- La metodología de reconocimiento de señas planteada en este trabajo demostró ser robusta tanto para señas aisladas como para señas compuestas, sin embargo cabe destacar que esta presenta un mejor rendimiento para las señas compuestas; esto debido a que el número de clases distintas de señas es menor en las señas compuestas.

- La base de datos se grabó con cinco actores sordos, para quienes la lengua de señas es su lengua nativa, y un actor oyente, quien es intérprete de lengua de señas. Existen marcadas diferencias en la forma de realización de las señas entre los actores sordos y el intérprete, las cuales se reflejan en la naturalidad del lenguaje. Los experimentos desarrollados en este trabajo demostraron que la metodología propuesta puede llegar a ser robusta en cuanto no se observan diferencias significativas en los resultados de clasificación entre las personas nativas y el intérprete.

6.2. Sistema de traducción

- En este trabajo se implementó un sistema de traducción automática basado en modelos estadísticos. Para el entrenamiento y validación de este sistema se construyó un corpus paralelo de 517 frases en español y glosas en lengua de señas en el dominio de aplicación específico seleccionado para este trabajo. Para la construcción de este corpus paralelo se contó con la asesoría de un intérprete experto en lengua de señas.

- Se entrenaron modelos probabilísticos de lenguaje basados en bigramas y trigramas a partir de las frases en español y las glosas del corpus paralelo.

- Se entrenó un modelo estadístico de traducción y se validó estadísticamente el rendimiento de la metodología propuesta mediante métricas empleadas en el campo del procesamiento del lenguaje natural tales como BLEU y WER.

- El corpus paralelo creado en este trabajo no contiene múltiples anotaciones de una misma frase, por tanto el sistema se valida contra una única frase objetivo. Por lo anterior se puede concluir que el empleo de la métrica WER es poco fiable a la hora de evaluar el rendimiento de este sistema de traducción automática, ya que esta medida depende del orden, número de palabras y ubicación de las mismas. Esta métrica busca la comparación exacta entre las frases evaluadas y no valida la traducción exitosa de frases semánticamente similares pero sintácticamente dispares.

- Se comprobó que no existen diferencias significativas entre los dos modelos creados para el sistema de traducción automática (modelo generado con bigramas y trigramas - ver figura 39 y 41); esto se debe a que si bien hay un mayor número de trigramas distintos, no existen las suficientes

repeticiones en la base de datos para que los modelos creados a partir de trigramas queden entrenados adecuadamente.

7. TRABAJOS FUTUROS

A continuación se mostrarán los trabajos futuros que se plantean una vez se realizó este documento

- Mejorar la forma de captura de la base de datos con el fin de tener un proceso más estándar y así eliminar condiciones propias de los autores.
- Realizar una fusión de sensores, utilizando IMU's o sensores de electromiografía, para así mejorar la eficiencia del sistema y poder reducir los errores que presenta el sistema de captura empleado.
- Crear un corpus paralelo con múltiples anotaciones, con el fin de optimizar el modelo de traducción automático y mejorar la eficiencia.
- Realizar un análisis semántico que permita plantear otras métricas para comprobar el modelo de traducción.

Referencias

[1] A. O. M. Tello, *Historia de La Comunicacion Humana (Spanish Edition)*. Palibrio, 2014. 1.2

[2] B. Comrie, Ed., *The World's Major Languages*. Routledge, 2009. 1.2

[3] M. Perez, "5 aplicaciones de traducción simultánea para tu smartphone," 2015, consultado: 2018-02-11. [Online]. Available: https://blogthinkbig.com/5-aplicaciones-de-traduccion-simultanea 1.2

[4] F. Vallejo, "Pilot, llega el traductor simultáneo, blog de tecnología e información," 2015, consultado: 2018-02-11. [Online]. Available: https://www.ibertronica.es/blog/actualidad/pilot-traductor-simultaneo/ 1.2

[5] J. Carbonell, "Cvc. congreso de sevilla. la lengua española y las nuevas tecnologías." 2015, consultado: 2018-02-11. [Online]. Available: https://cvc.cervantes.es/obref/congresos/sevilla/tecnologias/ponenc_carbonell 1.2

[6] B. Mundo, "Las 5 cosas menos conocidas sobre la sordera," 2018. [Online]. Available: http://www.bbc.com/mundo/noticias/2014/08/140808_salud_cinco_cosas_que_no_sabe_sordera_lv 1.2

[7] D. S. XXI, "Los sordos tienen dificultades para integrarse en su vida social y laboral," 2018. [Online]. Available: http://www.diariosigloxxi.com/texto-diario/mostrar/61445/los-sordos-tienen-dificultades-para-integrarse-en-su-vida-social-y-laboral 1.2

[8] A. Oviedo, "Apuntes para una gramática de la lengua de señas colombiana," República de Colombia, Ministerio de Educación Nacional, Instituto Nacional para Sordos, Tech. Rep., 2001. 1.2

[9] C. Savur and F. Sahin, "Real-time american sign language recognition system using surface emg signal," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, Dec 2015, pp. 497–502. 1.2, 2.2.1

[10] M. Georgi, C. Amma, and T. Schultz, "Recognizing hand and finger gestures with imu based motion and emg based muscle activity sensing," in *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 4*, ser. BIOSTEC 2015. Portugal: SCITEPRESS - Science and Technology Publications, Lda, 2015, pp. 99–108. [Online]. Available: https://doi.org/10.5220/0005276900990108 1.2, 2.2.1

[11] B. Plannerer, "An introduction to speech recognition," in *International Journal of Computer Applications*, vol. 12, no. 2, plannerer@ieee.org, 2005, pp. 1–7. 1.2

[12] G. M, M. C, U. P, and H. R, "A vision based dynamic gesture recognition of indian sign language on kinect based depth images," in *2013 International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA)*, Oct 2013, pp. 1–7. 1.2, 2.2.2

[13] M. Oszust and M. Wysocki, "Polish sign language words recognition with kinect," in *2013 6th International Conference on Human System Interactions (HSI)*, June 2013, pp. 219–226. 1.2, 2.2.2, 4.3.3, 5.1.1, 5.3

[14] H. V. Verma, E. Aggarwal, and S. Chandra, "Gesture recognition using kinect for sign language translation," in *2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*, Dec 2013, pp. 96–100. 1.2, 2.2.2, 4.3.3, 5.1.1, 5.1.2

[15] C. P. R. Rocha, J. A. P. Arias, and D. S. Villamil, "Prototipo traductor de señales manuales a texto legible, utilizando kinect," *Avances investigación en ingeniería*, vol. 10, no. 2, pp. 64–72, 2013. 1.2, 2.2.2, 5.1.1, 5.1.2, 5.3

[16] G. A. R. Fresneda, "Reconocimiento del lenguaje de señas manuales con el kinect," Master's thesis, Universidad de los Andes Colombia, Jan. 2014. 1.2, 2.2.2, 4.3.3, 5.1.1, 5.1.2, 5.3

[17] N. K. Caselli, Z. S. Sehyr, A. M. Cohen-Goldberg, and K. Emmorey, "ASL-LEX: A lexical database of american sign language," *Behavior Research Methods*, vol. 49, no. 2, pp. 784–801, may 2016. 2.1

[18] J. Piater, T. Hoyoux, W. Du, P. Dreuw, E. Efthimiou, T. Hanke, T. Johnston, E. M. Ruiz, and A. Schembri, "Proceedings of the 4th workshop on the representation and processing of sign languages: Corpora and sign language technologies," 2010. 2.1

[19] *The BSL SignBank Dictionary*, UCL Std. [Online]. Available: https://bslcorpusproject.org/data/ 2.1

[20] O. C. Julie A. Hochgesang, *ASL Signbank*, New Haven CT: Haskins Lab Std. [Online]. Available: https://aslsignbank.haskins.yale.edu/ 2.1

[21] C. for Sign Linguistics and D. Studies, *Asian SignBank*, Department of Linguistics and Modern Languages Std. [Online]. Available: http://cslds.org/asiansignbank/ 2.1

[22] T. Johnston:, *Auslan Signbank*, Australian Research Council Std. [Online]. Available: http://www.auslan.org.au/ 2.1

[23] R. Elliott, H. Cooper, J. Glauert, R. Bowden, and F. Lefebvre-Albaret, *Kinect Sign Recognition*, The Centre for Vision, Speech and Signal Processing, Std. [Online]. Available: http://cvssp.org/data/KinectSign/webpages/downloads.html. 2.1

[24] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, Dec. 2015. 2.1

[25] I. of ManâMachine Interaction, *Database for Signer-Independent Continuous Sign Language Recognition*, Institute of ManâMachine Interaction Std. [Online]. Available: http://www.phonetik.uni-muenchen.de/Bas/SIGNUM/ 2.1

[26] E. languages of the world, *Greek Sign Language*, Ethnologue languages of the world Std. [Online]. Available: https://www.ethnologue.com/language/gss 2.1

[27] T. Kapuściński and D. Warchoł, "A suite of tools supporting data streams annotation and its use in experiments with hand gesture recognition," *Studia Informatica*, vol. 38, no. 4, 2017. 2.1

[28] T. Kapuscinski and P. Organisciak, "Handshape recognition using skeletal data," *Sensors*, vol. 18, no. 8, p. 2577, 2018. 2.1

[29] C. Neidle, A. Thangali, and S. Sclaroff, "Challenges in development of the american sign language lexicon video dataset (asllvd) corpus," in *Proc. 5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon*. Citeseer, 2012. 2.1

[30] Y. Li, *Kinect Sign Language*, Microsoft Research Asia Std. [Online]. Available: http://vipl.ict.ac.cn/homepage/ksl/home.html 2.1

[31] A. Nandy, S. Mondal, J. S. Prasad, P. Chakraborty, and G. Nandi, "Recognizing & interpreting indian sign language gesture for human robot interaction," in *Computer and Communication Technology (ICCCT), 2010 International Conference on*. IEEE, 2010, pp. 712–717. 2.1

[32] F. Ronchetti, F. Quiroga, C. Estrebou, L. Lanzarini, and A. Rosete, "Lsa64: A dataset of argentinian sign language," *XX II Congreso Argentino de Ciencias de la ComputaciÃ³n(CACIC)*, 2016. 2.1

[33] A. J. Porfirio, K. L. Wiggers, L. E. Oliveira, and D. Weingaertner, "Libras sign language hand configuration recognition based on 3d meshes," in 2013 IEEE International Conference on Systems, Man, and Cybernetics. *IEEE, 2013, pp. 1588–1593. 2.1*

[34] A. Madushanka, R. Senevirathne, L. Wijesekara, S. Arunatilake, and K. Sandaruwan, "Framework for sinhala sign language recognition and translation using a wearable armband," in Advances in ICT for Emerging Regions (ICTer), 2016 Sixteenth International Conference on. *IEEE, 2016, pp. 49–57. 2.2.1*

[35] J. Wu, L. Sun, and R. Jafari, "A wearable system for recognizing american sign language in real-time using imu and surface emg sensors." IEEE J. Biomedical and Health Informatics*, vol. 20, no. 5, pp. 1281–1290, 2016. 2.2.1*

[36] S. Shin, Y. Baek, J. Lee, Y. Eun, and S. H. Son, "Korean sign language recognition using emg and imu sensors based on group-dependent nn models," in Computational Intelligence (SSCI), 2017 IEEE Symposium Series on. *IEEE, 2017, pp. 1–7. 2.2.1*

[37] N. Pugeault and R. Bowden, "Spelling it out: Real-time asl fingerspelling recognition," in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Nov 2011, pp. 1114–1119. 2.2.2*

[38] T. Shanableh and K. Assaleh, "Two tier feature extractions for recognition of isolated arabic sign language using fisher's linear discriminants," in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 2, April 2007, pp. II–501–II–504. 2.2.2*

[39] M. Mohandes, S. I. Quadri, and M. Deriche, "Arabic sign language recognition an image-based approach," in Advanced Information Networking and Applications Workshops, 2007, AINAW '07. 21st International Conference on*, vol. 1, May 2007, pp. 272–276. 2.2.2*

[40] S. Huang, C. Mao, J. Tao, and Z. Ye, "A novel chinese sign language recognition method based on keyframe-centered clips," IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 442–446, March 2018. 2.2.2*

[41] D. C. L. Bustamante and E. M. Arcila, Implementación de un sistema de traducción automática basado en modelos estadísticos para la traducción de la lengua de señas colombiana al español*, Universidad Tecnologíca de Pereira Std., 2017. 2.3, 5.3*

[42] A. Oviedo, Las señas tienen partes*, Instituto Caro y Cuervo Std., 2001. 3.1*

[43] J. Wu, Advances in K-means Clustering: A Data Mining Thinking (Springer Theses: Recognizing Outstanding Ph.D. Research). *Springer, 2012. 3.2.1*

[44] S. G. Rao and A. Govardhan, "Performance validation of the modified k-means clustering algorithm clusters data," International Journal of Scientific & Engineering Research*, vol. 6, no. 10, pp. 726–730, 2015. 3.2.2*

[45] X. D. Huang, "Phoneme classification using semicontinuous hidden markov models," IEEE Transactions on Signal Processing*, vol. 40, no. 5, pp. 1062–1067, May 1992. 3.3*

[46] C. E. Shannon, "A mathematical theory of communication," ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001. 3.4.3*

[47] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," Computational linguistics*, vol. 19, no. 2, pp. 263–311, 1993. 3.4.3*

[48] V. L. L. na, "Diseño, desarrollo y evaluaciónn de sistemas de mt para reducir las barreras de comunicación de las personas sordas," Ph.D. dissertation, Universidad Politécnica de Madrid, 2014. 3.4.3, 4.3.3, 5.2.2, 5.3

[49] P. Koehn, Statistical Machine Translation. *Cambridge University Press, 2009. 3.4.3*

[50] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting on association for computational linguistics. *Association for Computational Linguistics, 2002, pp. 311–318. 3.5.1*

[51] A. C. Morris, V. Maier, and P. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition," in Eighth International Conference on Spoken Language Processing*, 2004. 3.5.2*

[52] Y. Zhu, Y. Zhao, and S.-C. Zhu, "Understanding tools: Task-oriented object modeling, learning and recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2855–2864. 4.2.3*

[53] G. García-Bautista, F. Trujillo-Romero, and S. O. Caballero-Morales, "Mexican sign language recognition using kinect and data time warping algorithm," in Electronics, Communications and Computers (CONIELECOMP), 2017 International Conference on. *IEEE, 2017, pp. 1–5. 5.1.2, 5.3*

[54] Z. A. Ansari and G. Harit, "Nearest neighbour classification of indian sign language gestures using kinect camera," Sadhana*, vol. 41, no. 2, pp. 161–182, 2016. 5.1.2*

[55] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens et al.*, "Moses: Open source toolkit for statistical machine translation," in* Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. *Association for Computational Linguistics, 2007, pp. 177–180. 5.2, 5.2.1*